Dottorato di Ricerca in
Linguistica Generale, Storica, Applicata, Computazionale e delle Lingue Moderne

2002-2004
L-LIN/01

Representation and Inference for Open-Domain Question Answering: Strength and Limits of two Italian Semantic Lexicon

2006

Relatori

Prof.ssa Nicoletta Calzolari Zamorani
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Pisa

Prof. Alessandro Lenci
Dipartimento di Linguistica T. Bolelli,
Università degli Studi di Pisa
Pisa

Candidato

Francesca Bertagna

Dipartimento di Linguistica T. Bolelli,
Università degli Studi di Pisa
Pisa

Questa ricerca è stata finanziata dal Consiglio Nazionale delle Ricerche (CNR) con una borsa di dottorato di tre anni ed è stata svolta in collaborazione con l'Istituto di Linguistica Computazionale del CNR di Pisa.

Ringraziamenti

Desidero ringraziare le molte persone che, a vario titolo, hanno contribuito alla ricerca descritta in queste pagine.
Ringrazio in primis Nicoletta Calzolari, per la serietà con cui mi ha affiancato durante la stesura di questa tesi e per lo stimolo, l'incoraggiamento e il supporto.
Un ringraziamento particolare va ad Irina Prodanof, per il sostegno disinteressato ed i preziosi consigli con i quali mi è stata vicina durante tutti gli anni di dottorato.
A mio marito, Simone, va il mio grazie più sentito, non solo per la pazienza e l'appoggio ma anche per l'aiuto fondamentale nella concreta implementazione di alcuni moduli del prototipo.
Ringrazio Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni e Vito Pirrelli per avermi concesso l'uso dell'intera catena di analisi della lingua italiana, senza la quale non avrei potuto sviluppare il prototipo frutto di questa ricerca.
Allo stesso modo vorrei ringraziare Giuseppe Attardi, Maria Simi e il loro gruppo di ricerca al Dipartimento di Informatica dell'Università di Pisa per avermi messo a disposizione il motore di ricerca da loro sviluppato nonché la loro esperienza nel campo dell'Open-Domain Question Answering.
Un grazie particolare va inoltre agli organizzatori del track di Question Answering del Cross-Language Evaluation Forum (tra gli altri Bernardo Magnini ed Alessandro Vallin), per la possibilità che mi è stata data di usare, come materiale per gli esperimenti, le collezioni di domande e di testi da loro create.
Sono inoltre in debito con tutti i colleghi dell'ILC che mi hanno pazientemente sostenuto in questi anni, in particolare con le persone che condividono con me interessi e progetti di ricerca, Monica Monachini e Claudia Soria.

Alla mia famiglia

Table of Contents

Figures

Tables

Introduction

This chapter introduces the motivations of the thesis, highlights the most important issues and presents the dissertation plan.

Work objectives, motivations and methodology

> *What remains an open-ended question is whether* […] *general purpose lexicons and ontologies are actually useful and usable when they are constructed independently of specific NLP applications. This will remain a controversial and unanswered question to be verified once these knowledge resources will become available.(Busa and Bouillon, 2001)*
>
> *[…] more systematic assessment of the needs of various NLP tasks, an area which deserves serious attenction. (Ide and Veronis, 1993)*

This dissertation is dedicated to the exploration of the role of lexico-semantic language resources in a Question Answering application for Italian. We briefly define here Question Answering (QA) as an application which allows the user to obtain concise, relevant answers from text collections in response to written questions (a more complex and detailed introduction will be dedicated to QA in the following chapters). One of the things that makes Question Answering such a challenging task is the necessity to go beyond the literal form of the query and of the answer: in the attempt to bridge the gap between the question and the candidate answer, the system has to "understand" natural language, handle some representation of the meaning of the two texts and perform textual inference by working on relevant, unstated information. One way to tackle this challenge is to resort to lexico-semantic language resources, which are supposed to provide an explicit and machine understandable representation of word meaning that can be exploited by intelligent agents as a source of knowledge for supporting inference.

Since last '80s, the availability of large-scale, lexico-semantic computational lexicons was precisely what a part of the community of computational linguists said was required in order to permit effective and robust natural language processing systems such as machine translation, question-answering, natural language front-ends etc. (Amsler, 1989, Boguraev, 1987, and Calzolari, 1988). Not only were computational lexicons intended to be the keystone for natural language technologies, but a line of research was also based on the conviction that lexico-semantic information was quite easily extractable from implicit knowledge sources, i.e. definitions in Machine Readable Dictionaries

(MRDs)[1] (Calzolari, 1984, Chodorow *et al*., 1985, Byrd *et al*., 1987, Boguraev and Briscoe 1987). In Amsler's position paper (Amsler, 1987) we read:

> *"For several years now I have been concerned with how artificial intelligence is going to build the substitute for human world knowledge needed in performing the task of text understanding. I continue to believe that the bulk of this knowledge will have to be derived from existing machine-readable texts produced as a byproduct of computer typesetting and word processing technologies which have overtaken the publishing industries."*

In 1993, Ide and Véronis (Ide and Véronis, 1993) provided a preliminary balance of the concrete results of this line of research, claiming that the conviction that large knowledge bases could be generated automatically from MRDs revealed itself to be a false expectation and that, in order to overcome the serious inconsistency of dictionary definitions, their construction would have required an important effort in terms of human intervention. Moreover (Ide and Véronis, 1993) recognized that:

> *"..the exact nature and kind of information required for various NLP tasks has not been fully explored. …It is difficult to draw a precise taxonomy in many cases…and yet humans easily understand sentences containing words for which the taxonomic relations are unclear. This suggest that the kind of precision NLP researchers have traditionally sought in knowledge bases may be unnecessary in some cases….it is clear that more consideration of the exact requirement of various NLP tasks needs to be done".*

More than ten years have passed since the conclusions of that paper (conclusions that are now discussed again by one of the authors in (Ide and Wilks, forthcoming)): many difficulties concerning the acquisition of lexical information have been overcome and the research field is still alive and kicking: many wide/medium-coverage computational lexicons are now available for dozens of languages, generated (semi-)automatically or completely by hand. 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). 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) etc.

But even if the research dedicated to lexicon and lexicon acquisition progressed in many ways, after more than ten years other balances are in the pipeline: even if computational lexicons have not revealed themselves to be "killer-resources" for NLP applications, they are partly exploited in existing systems (even non-commercial ones). However, only few types of applications have obtained

[1] Other approaches, considering the possibility of acquiring information directly from free text, began in the same years (Hearst, 1992 *inter alia*).

important benefits from the use of lexico semantic information. For example, one of the applications whose performance was expected to improve most with the use of semantic information was Information Retrieval (IR). But Knowledge-Based IR seems to obtain successful results only when applied to very specific and narrow domains (Sparck Jones, 1999 reports the results of Rada *et al*., 1989 and Monarch and Carbonell, 1987) while results in open-domain have been disappointing (Richardson and Smeaton, 1995) and IR has remained basically a coarse task. The TREC experience has demonstrated how difficult it is to obtain good results using query expansion (Voorhers, 1994) and even the actual usefulness of explicit word sense selection, for which linguistic analysis would be required, is far from obvious (Krovetz and Croft, 1991). In fact, coordination effect and redundancy seem sufficient to retrieve pertinent documents (Schütze and Pedersen, 1995, Lewis, 1991 but cf. also the results of the Senseval-3 Panel on WSD and applications, 2004[2]). In (Sparck Jones, 1999) we read that content-based information management has to be sought elsewhere: the aim of IR is to display information to the user about whole documents, while giving selected phrases may give more information about actual document content than matching terms or listing key words. These goals and methods lead us to other final applications, such as information extraction, summarization and question answering.

This is an appropriate starting point from which to investigate whether/to what extent the information encoded in LRs can be exploited to support exigent information management functions. The choice to use QA is mostly due to the fact that literature on QA (Hirshman and Gaizauskas, 2001, Harabagiu *et al*., 2000) shows that it is an application that can benefit from the use of lexical semantics. Moreover, QA also incorporates an IR module that can be enriched by means of consolidated techniques of lexical query expansion (recurring to LRs) allowing us to try out LRs (in particular ItalWordNet) in one of their "natural" tasks[3]. Testing activity in QA task can be conceived as a possible way to evaluate the heuristic and predictive value of word meaning as instantiated in various language resources. Question Answering can be considered a sort of *sand-box*, a controlled environment where the usefulness and appropriateness of lexical-semantic information available in Language Resources can be tested and evaluated in the light of the requirements of a real application scenario. We think applications can highlight the potentialities, together with problems and limits, of the bulk of information that an important part of the community of linguists and computational linguists collected during the last two decades.

By observing the way the application "interacts" with the lexico-semantic information in the resources, we will try to provide answers to a series of questions: what type of information can be successfully exploited? What information is present in some forms in language resources but cannot

[2] Presentations available at http://www.senseval.org/senseval3/panels.
[3] QA is interesting also for its potentialities in an industrial perspective, being a core application for many other technologies (such as smart agents, e-commerce solutions, access to on-line documentation etc.).

be exploited by the system, and why? What information would be useful but is completely missing? And, above all: when the system has to "answer a question", do the representational devices and the very content of the lexical items constitute an adequate source of knowledge with respect to lexical complexity?

Many entities play a role in the development of this line of reasoning: in the background, we have the human being, a natural "cognitive agent" that interprets reality, interacts with the world and carries out complex tasks of a different nature (question answering but also climbing stairs or planning a holiday). Humans "grasp the world perceptually" (Jackendoff, 2002), alimenting the organization of conceptual information in the mind. As a lot of empirical evidence seems to suggest, cognitive processes are computational processes accomplished by operating on a large amount of information that has to be structured in some way in order to be accessible and useful (Caramazza and Shapiro, forthcoming). Language is an important mirror of these competences and provides clues about the way conceptual objects work in our brain. The information on language use that is present in the human *functional mind* (Jackendoff, 2002) is, in some ways, reproduced by lexico-semantic resources, which instantiate hypothesis on meaning representation, lexical access and language production. With many terminological or notational variations in its instantiation, the basis for most systems in computational linguistics was knowledge representation, i.e. the effort to represent knowledge about the world by means of organization of concepts, ontologies of types, structures of the type *genus* plus *differentiae*, selectional constraints in the possible concept combinations, semantic relations etc.

The design of the semantic network WordNet (Miller, 1985, Fellbaum, 1998) follows psycholinguistic principles on human memory and lexical access, FrameNet (Baker *et al*. 1998) proposes the Frame, an extended and complex structure of knowledge, as fundamental representational unit, the SIMPLE-CLIPS lexicons (Lenci *et al*. 2000, Ruimy *et al*. 2003) concretely encode Pustejovsky's Qualia Roles, etc.

All these resources were not conceived to meet the requirements of a specific task but rather to represent a sort of repository of information of general interest. A consequence of this *generalistic policy* is that, from the beginning, LRs have been built in a sort of aprioristic way with respect to applications, without actually considering the real use of the information encoded but rather recurring to traditional sources of lexical information, such as dictionaries, to select the information to encode for each lexical item and organizing that same information on the basis of a specific linguistic model.

But language resources are not only an attempt to encode linguistic information following particular theoretical principles: as conceptual objects contribute to functional activities in the human brain, language resources should allow systems to automatically perform inferences, distinguish senses, retrieve information, summarize texts, translate words in context from a language to another etc. The parallelism between human performance and applications cannot be taken completely for

granted: in some ways, it presupposes that having access to a knowledge-base isomorphic to the organization of concepts in the human brain is the solution for automatic natural language understanding. There are at least two objections that can be raised against this assumption: first of all, all we have about lexical organization in human mind are just hypotheses, partial results and something that is far from being uncontroversial and definitive. We do not have the perfect knowledge of how things work in our brain and we cannot simply reproduce the mechanism to see if it works in automatic processing as well. Moreover, it is not sure that exactly determining the modalities of lexical organization in the human brain would be the final solution for natural language processing: in theory, effective applications may also obtain a performance comparable to human performance by working on a completely different basis. A well-known and now historical case is represented by ELIZA (Weizenbaum, 1966), an early natural language processing system mimicking a Rogerian psychotherapist in a conversation with a human. ELIZA worked by exploiting only simple pattern-matching rules without *knowing absolutely anything* about the world (Jurafsky and Martin, 2000). But what is clear today is that we can expect to create applications that go beyond a simple mimicking of human intelligence and for those aims the possibility of exploiting information of a semantic nature seems, in theory, very promising.

We would like to assume an empirical attitude: we will firstly try to understand what type of information makes a text meaningful to people in the specific task of answering a question; with the help of a questionnaire, we will try to verify what types of inference are activated when a human recognizes in a text portion a plausible answer to a given question. Then, we will verify whether/to what extent the information already available in language resources can be used to support these types of inferences. This means that we expect to receive some evidence from the point of view of the nature of the information required to perform a QA task. For example, we will see how we should perhaps get over the traditional distinction between encyclopaedic versus lexical information if we want our systems to be capable of dealing with inference mechanisms in practical reasoning[4]. The next step consists in the attempt to exploit these information types in a real Question Answering prototype: this task represents an important part of overall work since no already existing applications were available for Italian when the research began. Many efforts were hence dedicated to the building of a prototype which constitutes an experimental environment where it has been possible to verify: i) what types of lexico-semantic information could be exploited, ii) which information is present in some forms in

[4] (Jackendoff, 2002) reports the same need to revise the distinction between encyclopaedic and lexical information. He reports the difficulties, for Bonnie Webber's research group, to program a virtual robot to respond to the natural language command "remove": the robot had to know what to do when told to remove something, but removing wallpaper from a wall requires a different action than removing a lid from a jar or a jar from a refrigerator etc. Jackendoff asks where such knowledge should be classified, as the encyclopaedic meaning of "remove" or as encyclopaedic meaning of wallpaper, lid or jar or as general world knowledge or somewhere else. Identifying this experience with our situation, we should ask ourselves whether this kind of information has to be present in a knowledge base or not.

language resources but cannot be exploited by the system, iii) which information would be useful but is completely or partially missing.

We would like to shed lights on the usefulness of computational lexicons in natural language processing, providing evidence from the point of view of a particular application. Now that many wide-coverage lexicons are available for dozens of languages, it is particularly important to verify whether they are up to the many natural language processing tasks they are born for or whether they can be modified and integrated to better support application requirements. Under the methodological point of view, we have to highlight the importance of exploiting a real application to support our work: not all the considerations that will arise from the actual use of the lexicons in application would have come to the surface on a theoretical level. As a matter of fact, computational processes have specific requirements in terms of characteristics of the representation in use: for example, the presence of given information in the knowledge-source is not sufficient to exploit this information in an application environment but the information has to be represented in a completely explicit, non ambiguous and systematic way. Even if representational issues are surely important, in the dissertation we will also investigate the adequateness of the *content* of language resources with respect to the task at hand. For example, we will try to determine the usefulness of a given semantic relation with the aim of identifying parts of the computational lexicon that should be boosted in order to increase the performance of the system. We will analyse the system performance starting from the evaluation of the results obtained by exploiting only non semantic modules (standard Information Retrieval techniques and syntactic parsing). These results will be considered a baseline that we will try to enhance recurring to lexico-semantic feedbacks. Obviously, the most interesting aspect will be the analysis of the system failures due to deficiencies or limits of language resources.

Notwithstanding this, the choice to use an application to evaluate language resources can be subjected to an objection: we said that we consider Question Answering a sort of controlled environment, a "place" where the variable *word meaning* can be observed while exploited to support different types of operations and inferences. It's obvious that Question Answering is not a true *constant* of the problem: a persistent, determined once-in-a-lifetime QA architecture does not exist; on the contrary, every year, in the event of the TREC and CLEF international competitions[5], we can observe the improvements of the performance of systems developed by exploiting the most different techniques and approaches. Moreover, as was said, at the beginning of the research no ready QA applications were available for Italian so, in some ways, the prototype has grown together with the

[5] The Text REtrieval Conference (TREC) is a series of workshops organized by the National Institute of Standards and Technology (NIST). The Cross-Language Evaluation Forum (CLEF) consists in annual evaluation campaigns and workshops offering the mono and multilingual QA track among a series of tracks designed to test different aspects of mono- and cross-language system performance.

Ph.D. work. From a methodological perspective, it could seem that we are dealing with a *two variables* (i.e. resources and application) problem. The specific solutions adopted in the system, the individual performance of the many modules of analysis making up the overall architecture, the very prototypical nature of the system and the consequent low level of engineering: all represent factors whose importance and weight have to be taken into account when evaluating the obtained results. Notwithstanding this, we do not really consider the variability of the application a point of weakness of our thesis. There are differences between the computational study of the lexicon and more traditional linguistic approaches and one of these differences is just the necessity to evaluate *utility*: how useful are the lexical entries for specific tasks and applications? What is the heuristic value of the lexical entries with respect to a specific task? The new approaches to modelling the structure of the lexicon recently emerged in computational linguistics (i.e. theoretical studies of how computations take place in the mental lexicon and developments of computational models of information in lexical databases) are somehow intertwined since the use of explicit representations can reveal the limitations of a given analytical framework. During our analysis, we will try to take all the factors concerning the choice of implementation into account, trying, at the same time, to keep the topic of the investigation well focused on the language resources, content and representational issues. (Nirenburg and Raskin, 1996) proposes an interesting distinction between two opposing methodological positions that can be detected in today's lexical semantics: the *Supply-Side* and the *Demand-Side*: the Supply-Side position belongs to research activities pursuing:

> *"the formulation of lexical meaning theories as algebraic entities in which the maximizing factor is formal elegance, descriptive power, economy of descriptive means, and absence of exceptions"* (Nirenburg and Raskin, 1996).

On the contrary, Demand-Side position can be characterized by the pursuing of theories which are capable of supporting practical applications. (Nirenburg and Raskin, 1996) also reports the description of a similar distinction made by (Wilks, 1994):

> *"There is a great difference between linguistic theory in Chomsky's sense, as motivated entirely by the need to explain, and theories, whether linguistic, AI or whatever, as the basis of procedural, application-orientated accounts of language. The latter stress testability, procedures, coverage, recovery from error, non-standard language, metaphor, textual content, and the interface to general knowledge structures".*

At a first glance, this dissertation could be thought as belonging to the Demand-Side position and, as a matter of fact, we will deal with issues (Nirenburg and Raskin, 1996) typically ascribing to this position, such as:

i) determining the number of lexemes in a lexicon (breadth),

ii) establishing criteria for sense specification and delimitation,

iii) granularity issue (determining the threshold of synonymy and ambiguity),

iv) tuning the depth and breadth of lexical description to the needs of a particular application.

However, we would also like to touch on an issue supposedly of interest to "supply-siders", i.e. formalism for representing lexical knowledge. As already said, these aspects are probably intertwined, because the outcomes of the demand-side research can provide feedback to the supply-siders.

Another difficulty that has to be kept in mind is that the object of the study can be considered somehow ambiguous: what do we mean when we say that we are going to evaluate a semantic lexicon? A lexicon is a bulk of lexical entries which i) follows a specific theoretical framework (psycholinguistic principles determined studying human lexical memory, type feature structures, lexical conceptual structures, conceptual frames, semantic networks, theory of shared semantic information based on orthogonal typed inheritance principles etc.) ii) adopts a given set of representational devices at low level (it can be a database, an XML file, a text file following particular in-house format), iii) generally results from the work of human encoders that, in a very human way, make mistakes and subjective choices. Each interaction between the QA system and the lexicon, either successful or unsuccessful, will be evaluated in the light of all these aspects but trying, when possible, to generalize.

Thus we will try to understand when a failure of the application is due to a mere lack of specific information or to a deficiency motivated by more general reasons. In this sense, we will consider in a different way the case of a missing derivational relation between a noun and a verb in a knowledge source that foresees the encoding of relations between verbs and deverbatives (a contingent "error") and the case of a pervasive impossibility of finding the required similarity relation between two concepts in a lexicon whose design is based on synonymy encoding. We will always try to distinguish and evaluate the different clues derived from the analysis of the exploitation of the lexicon content, in the attempt to individuate some generalizations. As a matter of fact, there is one thing that all the lexicons have in common: they all adhere to a very well-established modality of account of meaning, i.e. the symbolic one, in which the semantics of a lexical item is conveyed by its coordinates within a generally completely context independent system of symbols. What we want to investigate is whether this system of symbols is adequate for the endless challenges arising when an automatic procedure deals with natural language complexity. We will try to keep in mind the five questions raised in the dialogue on the nature of symbols, language and representations (Nirenburg and Wilks, 1999):

> *"Are representation languages natural language in any respect? Are languages (natural or representational) necessarily extensible as to sense? Are language acquisition and extensibility linked? If automatic acquisition is possible, what are the consequences for any*

representation/natural language? What are the consequences of all of this, if any, for representations for humans (versus for machines)?"

Another important issue is represented by the evaluation of concurrent approaches or techniques: if LRs provide any contribution to the successful question answering, is there any other way to obtain the same results without using LRs? And, in that case, are the alternatives to LRs more easy and robust to use?

At the Istituto di Linguistica Computazionale of the CNR two lexical resources have been built during the past years (and are still object of refinement) and are now available for testing: the Italian component of the EuroWordNet project, ItalWordNet (Roventini *et al.*, 2003) and the Italian lexicon belonging to the SIMPLE family, SIMPLE-CLIPS (Ruimy *et al.*, 2003). We decided to work on Italian not only because there is the general need to access and manage the content of an increasing number of web pages and information in languages other than English, but also because this is the first time these two resources can be tested and evaluated in an applicative environment. A first evaluation of these Italian resources (also for comparative aims and in view of a future, possible merge of the content of the two lexicons) will be hence a sort of by-product of this dissertation. The CLEF campaign, moreover, fosters the research on information retrieval and *content technology* solutions for languages different from English and the present research has benefited, from the participation in the QA track of the 2004 edition of the competition (Magnini *et al.*, 2004), an important chance to finalize a first version of the system, working on a controlled set of questions and answer pairs and on a common reference corpus of news and articles. The CLEF QA track represented an important exercise in individuating the most important problems, in discussing and studying possible solutions and also in sharing our first results in a collaborative and experimental environment. The experience gained will surely be of great importance in the further development of our work.

Dissertation plan

The first two chapters will introduce the two focuses of this dissertation, i.e. the lexico-semantic language resources and the Question Answering application. In the first chapter we will describe the two language resources under analysis, ItalWordNet (Roventini *et al.*, 2003) and SIMPLE-CLIPS (Ruimy *et al.*, 2003), their linguistic design an the type of lexical information that is there represented.

The second chapter is instead dedicated to QA and we will introduce its "history", the various dimensions that play a role in its implementation and in particular the "places" where language resources can be exploited.

In the third chapter, we will analyse the results of a questionnaire in the attempt to understand what types of information make a text meaningful to people in the specific task of answering a

question. We will try to individuate the types of inference that are activated when a human recognizes in a text portion a plausible answer to a question. Then we will verify whether/to what extent the information already available in language resources can be used to support those types of inference.

In the fourth chapter, we will present the construction schema of the QA prototype we built. In particular, we will verify whether the semantic information highlighted in the previous steps can be exploited in a real QA prototype.

The last chapter will be dedicated to a final analysis of the results of the prototype with the aim of discussing conclusions and open issues. In particular, we will analyse the results obtained by the prototype on the CLEF2004 test bench, highlighting both successful exploitation of the information stored in language resources and the problems encountered. In particular, we will try to provide an answer to the system failures when language resources fail in supporting system reasoning capabilities.

1

In this chapter, we introduce the two lexicons under analysis, their linguistic design and the type of information they store.

1.1 Lexical Knowledge Bases

(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, but the term can also be extended to software and tools that work on such resources. The subset of linguistic resources that is the focal point of this dissertation is the one which comprehends *computational semantic lexicons*, where lexical knowledge is expressed in terms of semantic relations, classification hierarchies, selectional patterns, case frames etc. (Grishman and Calzolari, 1997).

A survey of existing lexicons is not among the aims of this dissertation and the interested reader can refer to (Grishman and Calzolari, 1997; Sanfilippo *et al*., 1999; Calzolari *et al*., 2002) for an overview of the most important lexicons, their design and the type and quantity of information they store. We will instead look at two models of organization of semantic information, i.e. the ones that are instantiated i) in the lexicons belonging to the WordNet family (and more specifically in the Italian ItalWordNet database) and ii) in the SIMPLE-CLIPS lexicons. These two models represent what can be called a *lexical knowledge base* (LKB), that we can define by using Amsler's words:

> *A lexical knowledge base is a repository of computational information about concepts intended to be generally useful in many application areas including computational linguistics, artificial intelligence, and information science. It contains information derived from machine-readable dictionaries, the full text of reference books, the results of statistical analyses of text usages, and data manually obtained from human world knowledge. A lexical knowledge base is not intended to serve any one application, but to be a general repository of knowledge about lexical concepts and their relationships.* (Amsler, 1984)

Even if many researchers have tried to introduce more fine-grained distinctions in the terminology used in the field to refer to this kind of repositories (Boguraev and Levin, 1993 *inter alia*), we think that it is not misleading to state that the terms (*semantic*) *computational lexicon*, *lexical* (*semantic*) *database*, *semantic resources* and (*lexical*) *knowledge base* are basically used as interchangeable expressions in the literature. But the "universe" of computational lexicons comprehends different types of resources: for example, in (Sanfilippo *et al*., 1999) a rough classification is introduced, defining traditional Machine Readable Dictionaries, wordnets, taxonomies, dictionaries with features classifications, lexicons for Machine Translation, Higher Level Ontologies, traditional bilingual dictionaries.

Even if we will not provide a survey of the current approaches and existing lexicons, we think it may be useful to provide a (non-exhaustive) list of the major phenomena and information types that we can find

represented in this kind of resources. 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). Given the scope of the current discussion, 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

In the following report, we will introduce the linguistic design of two lexical resources that are available for Italian and that we think are very representative of computational lexicons.

The ItalWordNet database

ItalWordNet is the extension of the Italian component of the EuroWordNet database (Ide *et al*., 1998). Both ItalWordNet and EuroWordNet are based on a common underlying linguistic design: the WordNet database.

1.1.1 WordNet

WordNet (Fellbaum, 1998) is a lexical database which contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of *synset*. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, {car; auto; automobile; machine; motorcar} form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss: "4-wheeled; usually propelled by an internal combustion engine". Finally, synsets can

be related to each other by semantic relations, such as hyponymy (between specific and more general concepts), meronymy (between parts and wholes), cause, entail, pertains, attribute_of, antonymy.

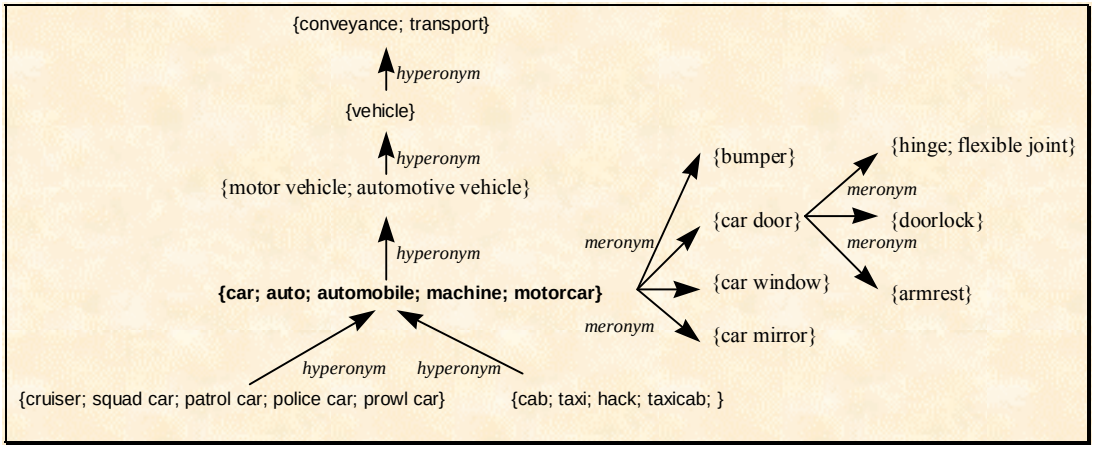


Fig. 1: Synsets related to "*car*" in its first sense in WordNet1.5.

In the example of Fig. 1 , taken from WordNet1.5, the synset {car; auto; automobile; machine; motorcar} is related to:

- a more general concept or the hyperonym synset: {motor vehicle; automotive vehicle},
- more specific concepts or hyponym synsets: e.g. {cruiser; squad car; patrol car; police car; prowl car} and {cab; taxi; hack; taxicab},
- parts it is composed of: e.g. {bumper}; {car door}, {car mirror} and {car window}.

Each of these synsets is again related to other synsets as is illustrated for {motor vehicle; automotive vehicle} that is related to {vehicle}, and {car door} that is related to other parts: {hinge; flexible joint}, {armrest}, {doorlock}. By means of these and other semantic/conceptual relations, all word meanings in a language can be interconnected, constituting a wordnet. Such a wordnet can be used for making semantic inferences (what things can be used as *vehicles*), for finding alternative expressions or wordings (what words can refer to *vehicles*), or for simply expanding words to sets of semantically related or close words, in e.g. information retrieval. Furthermore, semantic networks give information on the lexicalization patterns of languages, on the conceptual density of areas of the vocabulary and on the distribution of semantic distinctions or relations over different areas of the vocabulary. In (Fellbaum, 1998) a detailed description is given of the history, background and characteristics of the Princeton WordNet.

WordNet is fundamentally different from other computational lexicons because the semantic information is mainly stored for *synsets*, considered as conceptual units, rather than for word senses (and we will see that this is the predominant difference between WN and SIMPLE-CLIPS).

1.1.1.1 EUROWORDNET

The main goal of the EuroWordNet (EWN) project[6] was to develop a (multilingual) lexical resource, retaining the basic underlying design of WordNet, at the same time trying to improve it in order to answer the needs of research in the field of NLP. A fundamental change made in EWN was that the set of lexical relations to be encoded between word meanings was extended or modified in various ways with respect to the set defined in WN1.5 (Vossen, 1998).

Semantic information was encoded, within EWN, for about 50,000 word senses (nouns and verbs) in each of the languages treated, in the form of lexical semantic relations between *synsets* (i.e. synonym sets, cf. section 3). A rich framework of relations was designed which were considered useful for computational applications, for example the near_synonymy relation among different parts of speech. This decision was motivated by the requirements of a number of potential applications of EWN (most of all information retrieval) where it is essential to have a link between different lexicalizations (possibly through different parts of speech) with the same underlying meaning.

Within the framework of a National Project, SI-TAL[7], the Italian wordnet built in EWN was enlarged and improved. In the next section, we describe the overall architecture of the IWN database. It will be possible to see that a set of language-independent modules was foreseen in order to build an architecture that would be fully exploitable in cross-language tasks. We will refer to such language independent modules since they are functional to the information flow among modules but we will not describe them in detail since the scope of the current dissertation is limited to the monolingual QA. Moreover, we did not add the terminological modules dedicated to specific technical fields such as *law* and *finance* since they are not important for Open-Domain QA. On the contrary, particular emphasis will be dedicated to on the relations encoded since they are the type of information more suitable to support inferences.

1.1.1.2 OVERALL ARCHITECTURE OF THE IWN DATABASE

The IWN database is made up of the following components:

- a generic wordnet, built by extending the network developed within EWN, which contains about 46,000 lemmas corresponding to roughly 49,000 synsets and 65,000 word-senses;
- an Interlingual-Index (ILI) which is an unstructured version of WN1.5, containing all the synsets found in WN1.5 but not the relations among them. This module was used in EWN to link wordnets of different languages. Also in IWN the Italian synsets are linked to this interlingual index, to make the resource usable in multilingual applications;

[6] EWN (http://www.illc.uva.nl/EuroWordNet/) was a project in the EC Language Engineering (LE-4003 and LE-8328) programme.

[7] SI-TAL (Integrated System for the Automatic Treatment of Language) was a National Project devoted to the creation of large linguistic resources and software tools for Italian written and spoken language processing. Besides IWN, within the project were developed: a treebank with a three-level syntactic and semantic annotation, a system for integrating NL processors in applications for managing grammatical resources, a dialogues annotated corpus for applications of advanced vocal interfaces, software and tools for advanced vocal interfaces.

- the Top Ontology (TO), a hierarchy of about 60 language-independent concepts, reflecting fundamental semantic distinctions, built within EWN and partially modified in IWN to account for adjectives. The TO consists of language-independent features which may (or may not) be lexicalized in various ways, or according to different patterns, in different languages (Rodriguez *et al.*, 1998). Via the ILIs, all the concepts in the generic and specific wordnets are directly or indirectly linked to the TO;

All these components and their reciprocal links, i.e. the IWN architecture, are shown in Fig. 2.

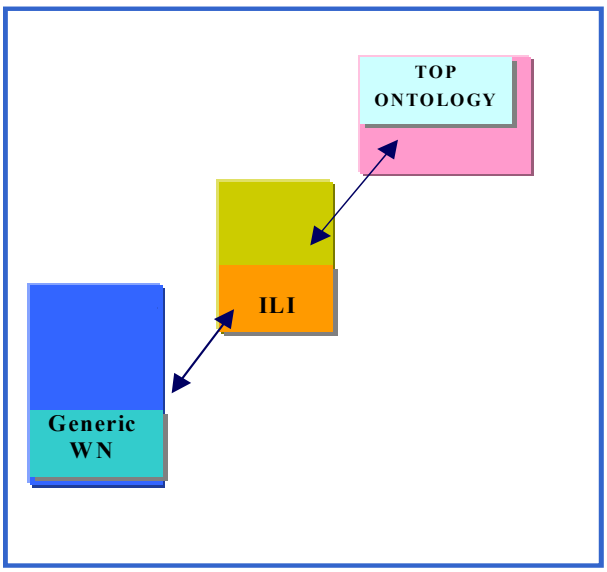


Fig. 2: the overall architecture of the (Euro/Ital)WordNet database

1.1.1.3 THE IWN LINGUISTIC MODEL

The basic notion around which the IWN database is built is the same around which both WN and EWN are built, i.e. that of a *synset* or set of synonymous words belonging to the same Part-of-Speech (PoS) that can be interchanged at least in a context. The notion of synonymy adopted, thus, is not the strongest one, which maintains that two expressions are synonymous if the substitution of one with the other never changes the truth value of a sentence in which the substitution is made. Instead, a weaker definition is adopted stating that "two expressions are synonymous in a linguistic context C if the substitution of one with the other in C does not alter the truth value" (Miller *et al.*, 1990). One such context is sufficient to state a synonymy relation between word senses, on the basis of which a synset is built. Within a synset we only find word senses, or multiwords or also acronyms, of the same PoS, called *variants* of the synset[8].

Synsets are linked mainly through the hyponymy (or IS-A) relation, but various other relations were encoded, partly inherited from EWN, to better describe the semantic relations among the synsets. In particular, the set of relations encoded in WN was enriched, both in EWN and in IWN, with relations

[8] Note that a synset may sometimes contain one word (sense) alone, if no synonyms are found for that word (sense). A synset is indicated by using braces, e.g. {hot, warm}.

applying between synsets belonging to different PoSs. In WN each PoS forms a separate network of language-internal relations, and therefore conceptually close concepts are clearly separated only because they differ in PoS. For instance, no relation links the noun *adornment* and the verb *to adorn,* although they refer to the same process ("the act of decorating oneself with something").

To avoid this separation between PoSs, which were traditionally identified by using a mixture of morphological, syntactic and semantic criteria, in EWN a distinction was drawn among the semantic orders of the entities to which word meanings refer (Lyons, 1977): 1st order entities (referred to by concrete nouns), 2nd order entities (referred to by verbs, adjectives or nouns indicating properties, states, processes or events), and 3rd order entities (referred to by abstract nouns indicating propositions independent of time and space)9. On the basis of this distinction, in IWN, as in EWN, various relations applying across PoSs were encoded. This approach does not seem merely more appropriate from a theoretical point of view, given that the distinction drawn is clearly based on semantic grounds, but can yield remarkable advantages with respect to the use of the database both for Information Retrieval purposes and for other Language Engineering applications (Alonge *et al*., 1998; Gonzalo *et al*., 1998).

IWN also inherited from EWN the distinction between *language-internal relations* and *equivalence relations*. The former link the language-specific synsets, while the latter link the Italian synsets to the ILI. By linking IWN to the ILI the possibility of use IWN for multilingual applications was ensured. IWN inherited the EWN language-internal relations (and related tests) with some minor changes. In the following a description of all the language-internal relations encoded in IWN is provided.

1.1.1.3.1 Internal relations

In order to encode relations in a consistent way substitution tests or diagnostic frames based on 'normality judgements' (Cruse, 1986) were used. Inserting two expressions in the same frame determines a 'semantic normality' judgement on the basis of which a relation can be detected.

Near_synonymy and xpos_near_synonymy

Within EWN it was observed that in some cases there is a close relation between synsets, resembling synonymy, which is not however sufficient to make them members of the same synset, i.e. they do not yield clear scores for the previous test or their hyponyms cannot be interchanged. For these synsets, as in EWN, the near_synonymy relation was used, which allows sets of hyponyms to be kept separate while encoding that two synsets are closer in meaning than other co-hyponyms. For instance, *disturbo* (disorder, upset) and *malattia* (disease, illness) have a very similar meaning, but they cannot share their respective hyponyms. Thus, they were linked by a near_synonymy relation. A similar relation also links synsets belonging to

[9] We do not think it is necessary to describe here in details the structure of the IWN Top Ontology. For a detailed description the reader is referred to (Roventini et al., 2003).

different PoSs. For instance a xpos_near_synonymy relation is encoded between the noun ricerca (research) and the verb ricercare (to research), which in fact indicate the same situation or eventuality.

Hyperonymy/hyponymy and xpos_hyperonymy/hyponymy

The hyperonymy/hyponymy relation corresponds to the class-inclusion logical relation and is an *asimmetric* and *transitive* relation.

The hyperonymy/hyponymy relation is the most important relation encoded for nouns and verbs both in WN and EWN, together with synonymy. This is due to the possibility it provides to identify classes of words for which it is possible to draw generalizations and inferences, e.g. fundamental semantic characteristics displayed by a node are inherited by all its sub-nodes.

While EWN contains detailed information only on nouns and verbs and therefore there are no hyponymy relations between adjectives or adverbs, the lack of such a relation for adjectives and adverbs in WN is mainly due to theoretical reasons. In WN adjectives are divided into two major classes: descriptive adjectives and relational adjectives. Typically, among the "descriptive" group, we find adjectives that designate the physical dimension of an object, its weight, abstract values, etc. Relational adjectives, on the other hand, mean something like "relating/pertaining to, associated with", and usually have a morphologically strong link with a noun. Typical examples are musical, atomic, and chemical. The organization of descriptive adjectives in WN can "be visualized in terms of barbell-like structures, with a direct antonym in the centre of each disk surrounded by its semantically similar adjectives (which constitute the indirect antonyms of the adjectives in the opposed disk)" (Fellbaum, 1998).

The main relation encoded for these adjective synsets is antonymy, claimed to be the most prominent relation, both from a psycholinguistic and a more strictly lexical-semantic point of view, in the definition of the semantics of descriptive adjectives. Hyponymy is substituted by a 'similarity' relation. Relational adjectives are not organized in the same way as descriptive adjectives because their semantics cannot be described by using antonymy and similarity relations. Indeed, they only point to the noun to which they pertain (e.g. atomic is linked to atom).

Although we also consider antonymy as the basic relation to define the semantics of most descriptive adjectives, we reconsidered the possibility of encoding hyponymy for this category. By analysing data from machine-readable dictionaries we found subsets of adjectives which have a *genus* + *differentia* definition, like nouns or verbs. These adjectives can be organised into classes sharing a superordinate. This is the case, e.g., of adjectives indicating a 'containing' property (*acquoso* - watery; *alcalino* - alkaline), or a 'suitable-for' property (*difensivo* - defensive; *educativo* - educational), etc. Hyponymy was thus also encoded for these sets of adjectives.

The hyponymy relation has also been encoded across PoSs: e.g., *entrata* (entering) is linked to *andare* (to go) by means of a xpos_has_hyperonym relation.

Antonymy and xpos_antonymy

In EWN two antonymy (semantic opposition) relations are encoded, namely antonymy, expressing meaning opposition between variants (used for cases in which it is not clear whether the opposition between two words may be extended to the synsets containing them), and near_antonymy, expressing synset oppositions. In IWN we assumed that since a synset contains different expressions for the same concept, it should not be possible to find an antonym of one of such expressions which is not antonym of the others. Thus, antonymy was only allowed between synsets. However, besides the general, underspecified antonymy relation we had the possibility of encode more specific sub-relations.

Following theoretical work (Lyons, 1977; Cruse, 1986), a further distinction between complementary_antonymy and gradable_antonymy was introduced. The former relation links synsets referring to opposing properties/concepts: when one holds the other is excluded (alive/dead). The latter relation is used for those antonym pairs which refer to gradable properties (long/short). In case it is not clear whether two opposing synsets refer to complementary or gradable concepts, we could still use an underspecified antonymy relation. This information can also be useful for computational applications since word pairs presenting one of the two kinds of opposition may occur in different contexts (Cruse, 1986).

Also the antonymy relation could be encoded across PoSs, by means of the xpos_antonymy relation: e.g., arrivo (arrival) is a xpos_antonym of partire (to leave).

Meronymy

In WN1.5 three kinds of part-of relations are distinguished (part/whole; member/group and component/substance). In IWN together with an underspecified relation (called HAS_MERONYM/HAS_HOLONYM) five sub-relations have been distinguished, also used in EWN:

- a whole-part relation
  mano (hand) HAS_MERO_PART dito (finger, toe)
  dito HAS_HOLO_PART mano
  piede (foot)

- a set-member relation
  senato (senate) HAS_MERO_MEMBER senatore (senator)
  senatore HAS_HOLO_MEMBER senato

- an object-substance relation
  muro (wall) HAS_MERO_MADEOF cemento (cement)
  cemento HAS_HOLO_MADEOF muro

- a whole-portion relation
  pane (bread) HAS_MERO_PORTION fetta (slice)
  fetta HAS_HOLO_PORTION pane

- a relation between a place and another place contained in it
  deserto (desert) HAS_MERO_LOCATION oasi (oasis)
  oasi HAS_HOLO_LOCATION deserto.

This relation was only encoded for concrete nouns.

Cause relations

In WN the cause relation may only be used to link verbs. In IWN, as in EWN, this relation is used to connect different 2nd order entities. Furthermore, in IWN various sub-relations of the underspecified CAUSE relation were distinguished:

- RESULTS_IN:

uccidere (to kill)	RESULTS_IN	morire (to die)
rotto (broken)	IS_RESULT_OF	rompersi (to break)

- FOR_PURPOSE_OF:

cercare (to search)	FOR_PURPOSE_OF	trovare (to find)
riuscire (to succeed)	IS_PURPOSE_OF	tentare (to try)

- IS_MEANS_FOR:

calore (heat)	IS_MEANS_FOR	distillazione (distillation)
evaporare (to evaporate)	HAS_MEANS	calore (heat)

In IWN we also used these sub-relations to encode some new data on causality.

Subevent

In IWN the has_subevent relation between 2nd order synsets characterizes the reference to situations occurring during the same stretch of time, where one *situation* includes the other, i.e. has the other as a *sub-event*: for instance, *to snore* is a sub-event of *to sleep*.

russare (to snore)	IS_SUBEVENT_OF	dormire (to sleep)	
dormire	HAS_SUBEVENT	russare	*reversed*
comprare (to buy)	HAS_SUBEVENT	pagare (to pay)	
pagare	IS_SUBEVENT_OF	comprare	*reversed*

Involved/Role

The INVOLVED/ROLE relation was defined to encode information on arguments (1st or 3rd order entities) clearly *incorporated* (lexicalized) within the meaning of 2nd order entities. When the relation links a 2nd order to a 1st/3rd order entity it is called INVOLVED, vice versa it is called ROLE relation. Besides the underspecified relation (used for unclear cases of involvement), a number of specific sub-relations is available:

- a relation between a 2nd order entity and an *agent* typically implied in its meaning:

sgambettare (to kick one's legs about)	INVOLVED_AGENT	neonato (baby)
pedone (pedestrian)	ROLE_AGENT	camminare (to walk)

- a relation between a 2nd order entity and a *patient* implied in its meaning:

partorire (to deliver a baby)	INVOLVED_PATIENT	figlio (child)
alunno (student)	ROLE_PATIENT	insegnare (to teach)

- a relation between a 2nd order entity and an *instrument* implied in its meaning:

bastonare (to cane)	INVOLVED_INSTRUMENT	bastone (cane)
pistola (gun)	ROLE_INSTRUMENT	sparare (to shoot)

- a relation between a 2nd order entity and the *location* where the situation it refers to occurs:

nuotare (to swim)	INVOLVED_LOCATION	acqua (water)
scuola (school)	ROLE_LOCATION	insegnare (to teach)

- a relation between a 2nd order entity and the goal or source of the movement it refers to; this relation can be underspecified (direction unspecified) or further specified:

condurre (to lead)	INVOLVED_DIRECTION	luogo (place)	
luogo	ROLE_DIRECTION	condurre	*reversed*
sbarcare (to disembark)	INVOLVED_SOURCE_DIRECTION	nave (ship)	
fonte (spring)	ROLE_SOURCE_DIRECTION	scaturire (to spring)	
rincasare (to go back home)	INVOLVED_TARGET_DIRECTION	casa (home)	
traguardo (goal)	ROLE_TARGET_DIRECTION	gara (competition)	

- a relation between a 2nd order entity and the result it produces, when this result is referred to by a 1st order entity (otherwise a CAUSE relation is used):

ghiacciare (to freeze)	INVOLVED_RESULT	ghiaccio (ice)
vapore (steam)	ROLE_RESULT	evaporazione (evaporation)

Since the kind of arguments incorporated within the meaning of a 2nd order entity determine both its semantic preferences and syntactic behaviour, encoding of this relation allows the user to obtain information which can be very useful for computational applications.

Co_role

co_role relations were defined in EWN and also used in IWN to encode links between 1st order entities which have a role in the same situation: e.g., pianista (pianist) and pianoforte (piano) have a role in the situation referred to by suonare il pianoforte (to play the piano).

Be_in_state

The BE_IN_STATE/STATE_OF relation is used to indicate the link between a 1st order and a 2nd order entity expressing the state in which the former is: e.g. a *poor* is someone who is in the state of being *poor*:

povero (poor man) (N)	BE_IN_STATE	povero (poor) (A)
povero (poor man) (N)	BE_IN_STATE	povertà (poverty) (N)
povero (A)	STATE_OF	povero (poor man) (N)
povertà (N)	STATE_OF	povero (poor man) (N)

In_manner

The IN_MANNER/MANNER_OF relation is used to encode a link between 2nd order entities and adverbs or adverbial expressions indicating the way in which the eventuality referred to occurs, when the 2nd order entity clearly refers to this modality:

bisbigliare (to whisper)	IN_MANNER	a bassa voce (in a low voice)
a bassa voce (in a low voice)	MANNER_OF	bisbigliare (to whisper)

Pertains_to

The PERTAINS_TO relation allows the link of a noun and a relational adjective: e.g. *musicale/musica* (musical/music), *presidenziale/ presidente* (presidential/president), etc. Among relational adjectives we also find ethnical adjectives, by using this relation we also linked relational adjectives to the relative proper nouns:

italiano	PERTAINS_TO	Italia
Italia	HAS_PERTAINED	italiano
musicale	PERTAINS_TO	musica
musica	HAS_PERTAINED	musicale

Is_a_value_of

A relation used in WN links an adjective to the noun of which it expresses a 'value'. For instance, *tall* expresses a value of *stature*. This relation, in a few cases, was also encoded in IWN, since it could be useful both to distinguish between adjective senses and to point out the adjective semantic preferences:

alto (tall)	IS_A_VALUE_OF	{statura, altezza} (stature)
alto (high)	IS_A_VALUE_OF	altezza (height)

Derivation

The DERIVATION relation was used to encode derivation links when no other semantic relation is available and it connects variants belonging to different PoSs:

acqua (water)	DERIVATION	*acquaiolo* (water carrier/seller)
grande (great)	DERIVATION	*grandemente* (greatly)

Has_instance

The HAS_INSTANCE/BELONGS_TO_CLASS relation was used to link proper nouns to the class of common nouns to which they belong:

fiume (river)	HAS_INSTANCE	*Danubio*
Roma	BELONGS_TO_CLASS	*città* (city)

Liable_to

A LIABLE_TO relation has been defined in IWN to encode information on a large group of deverbal adjectives expressing the possibility for an eventuality to occur:

giudicabile (judgeable)	LIABLE_TO	giudicare (to judge)
giudicare	HAS_LIABILITY	giudicabile

Fuzzynym

The EWN FUZZYNYM relation was used for all those cases in which it was not clear what kind of semantic relation connects two synsets which we found (by applying specific test frames) to be linked as in the example below:

collaborazionista (collaborationist) FUZZYNYM *nemico* (enemy)

A XPOS_FUZZYNYM relation may also be encoded.

The following table provides an overview of the relations encoded in IWN. For each relation we indicate: i) the semantic order of the entities linked; ii) one or more examples (provided in English).

Relation	*Order*	*Examples*
SYNONYMY	1°/1°; 2°/2°; 3°/3°	Bicycle/bike To analyse/to examine
NEAR_SYNONYM	1°/1°; 2°/2°; 3°/3°	Implement/utensil Cerebration/opinion
XPOS_NEAR_SYNONYM	2°/2°	Arrival/to arrive
HAS_HYPERONYM/HAS_HYPONYM	1°/1°; 2°/2°; 3°/3	Dog/animal To move/to travel
HAS_XPOS_HYPERONYM/HAS_XPOS_HYPONYM	2°/2°	Arrival/to go To hit/knock
ANTONYM	1°/1°; 2°/2°; 3°/3	To arrive/to leave
COMPL_ANTONYM	1°/1°; 2°/2°; 3°/3	Alive/dead
GRAD_ANTONYM	1°/1°; 2°/2°; 3°/3	Cold/hot
XPOS_ANTONYM	2°/2°	arrival/departure
HAS_HOLONYM/ HAS_MERONYM	1°/1°	arm/body hand/finger
HAS_MERO_PART/ HAS_HOLO_PART	1°/1°	foot/toe hip/body
HAS_MERO_MEMBER/ HAS_HOLO_MEMBER	1°/1°	team/player student/school
HAS_MERO_MADEOF/HAS_HOLO_MADEOF	1°/1°	jam/fruit
HAS_MERO_PORTION/ HAS_HOLO_PORTION	1°/1°	bread/slice slice/cake
HAS_MERO_LOCATION/ HAS_HOLO_LOCATION	1°/1°	city/city-centre oasis/desert
CAUSES/ IS_CAUSED_BY	2°/2°	to kill/to die execute/sentence
RESULTS_IN/ IS_RESULT_OF	2°/2°	to kill/to die sick/to fall ill
FOR_PURPOSE_OF/ IS_PURPOSE_OF	2°/2°	to search/to find to win/to compete
IS_MEANS_FOR/ HAS_MEANS	2°/2°	heat/distillation to evaporate/boiling
HAS_SUBEVENT/ IS_SUBEVENT_OF	2°/2°	to buy/to pay to snore/to sleep
INVOLVED/ ROLE	2°/1° 1°/2°	to hammer/hammer pedestrian/to walk
INVOLVED_AGENT/ ROLE_AGENT	2°/1° 1°/2°	to teach/teacher runner/to run
INVOLVED_PATIENT/ROLE_PATIENT	2°/1° 1°/2°	to teach/student student/to teach
INVOLVED_INSTRUMENT/ROLE_INSTRUMENT	2°/1° 1°/2°	to paint/paint-brush gun/to shoot
INVOLVED_LOCATION/ROLE_LOCATION	2°/1° 1°/2°	to swim/water school/to teach
INVOLVED_DIRECTION/ROLE_DIRECTION	2°/1° 1°/2°	to lead/place arrival/to arrive
INVOLVED_SOURCE_DIRECTION/ ROLE_SOURCE_DIRECTION	2°/1° 1°/2°	to disembark/ship outside/to enter
INVOLVED_TARGET_DIRECTION ROLE_TARGET_DIRECTION	2°/1° 1°/2°	to exit/outside inside/to enter
INVOLVED_RESULT/ROLE_RESULT	2°/1° 1°/2°	To freeze/ice Ice/to ice
CO_ROLE	1°/1°	Piano player/piano
GENT_PATIENT/CO_PATIENT_AGENT	1°/1°	Teacher/student Student/teacher
CO_AGENT_INSTRUMENT/ CO_INSTRUMENT_AGENT	1°/1°	Guitar player/guitar Guitar/guitar player
CO_AGENT_RESULT/	1°/1°	Painter/painting

CO_RESULT_AGENT		Painting/painter
CO_PATIENT_INSTRUMENT/ CO_INSTRUMENT_PATIENT	1°/1°	Wood/axe Axe/wood
CO_PATIENT_RESULT/ CO_RESULT_PATIENT	1°/1°	Skin/scarification Scarification/skin
CO_INSTRUMENT_RESULT/ CO_RESULT_INSTRUMENT	1°/1°	Camera/photo Photo/camera
BE_IN_STATE/STATE_OF	1°/2° 2°/1°	Poor/poorness Oldness/old
IN_MANNER/IS_MANNER_FOR	2°/2°	To whisper/ In a low voice
DERIVATION	All	Water/water-carrier
LIABLE_TO/HAS_LIABILITY	2°/2°	Judgeable/to judge
IS_A_VALUE_OF/HAS_VALUE	2°/2°	Tall/stature
PERTAINS_TO/HAS_PERTAINED	2°/1°, 2°/pn	Presidential/president Italian/Italy
HAS_INSTANCE/BELONGS_TO_CLASS	1°/pn	River/Po Rome/city
FUZZYNYM	All	Collaborationist/enemy
XPOS_FUZZYNYM	2°/2°	To govern/ Government-in-exile

Table 1: lexico-semantic relations in IWN (from Roventini *et al*., 2003)

1.1.2 Semantic information in the SIMPLE-CLIPS database

The SIMPLE-CLIPS database (Lenci *et al*., 2000) was developed in the framework of SIMPLE[10], a project aimed at building wide-coverage, multipurpose and harmonised computational semantic lexica linked to the morphological and syntactic ones which were elaborated for 12 European languages[11], during the PAROLE project. The Italian component of the SIMPLE-CLIPS lexicons was further developed in a national project with the name of CLIPS semantic lexicon (Ruimy *et al*., 2003). In this dissertation we will thus always refer to this lexicon with the name of SIMPLE-CLIPS.

An extended version of Pustejovsky's Generative Lexicon (GL) (Pustejovsky, 1995) provides the theoretical linguistic background of this database. Pustejovsky defines the semantics of a lexical item recurring to the *qualia structure*, a rich and structured representation of the relational force of a lexical item. The importance of this structure is that it allows the overcoming of the one-dimensional inheritance that is captured via standard hyperonymic relations, enabling the expression of orthogonal aspects of word sense.

As a matter of fact, lexical entries are generally organized according to taxonomical relations since many word senses can be entirely characterized in terms of a hierarchical relation to other lexical units. However, a substantial amount of word senses denoting a more complex bundle of lexical orthogonal dimensions that cannot be exhaustively captured in terms of a mere hyperonymic relation. The qualia structure allows multidimensionality of meaning to be encoded by means of four qualia roles which express essential aspects of a word's meaning:

[10] A Language Engineering project funded by EC DGXIII, which started in 1998 as a follow-up of the PAROLE project and ran for twenty-four months.

[11] Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.

- the *formal role* identifies an entity among other entities, it indicates therefore its position within the ontology of types;
- the *constitutive role* expresses the entity's composition, its constituent elements;
- the *agentive role* provides information about its origin, its coming about;
- the *telic role* specifies its function.

In SIMPLE-CLIPS the lexical entry is constituted by the word sense and called *semantic units* (*SemU*) and the qualia roles have been implemented as relations between SemU. Subtypes has been assigned to the four qualia roles. The example reported in (Ruimy *et al*., 2003) concerns 'photographer' for which (in order to preserve the information that to be a photographer may be either a profession or a hobby), different telic subtype relations were encoded, namely 'is_the_activity_of' and 'is_the_ability_of'.

In the Extended Qualia structure the relevance of a relation is marked with a different weight, for each of its actual uses in a type definition. The weight indicates whether the relation is *type defining*, i.e. encoding an information that intrinsically characterizes a semantic type or whether it conveys 'optional' - mainly world-knowledge – information.

1.1.2.1 THE SIMPLE-CLIPS ONTOLOGY OF TYPES

In the SIMPLE-CLIPS lexicon, semantic units are classified according to the semantic type system, mappable on the EuroWordNet ontology and consisting of a set of 153 language-independent semantic types, which are of two different kinds: *SIMPLE-CLIPS* and *unified* types.

- *SIMPLE-CLIPS types* can be fully characterized in terms of a hyperonymic relation;
- *unified types* can only be identified through the combination of a subtyping relation and the reference to orthogonal (telic or agentive) dimensions of meanings.

Moreover, the SIMPLE-CLIPS ontology is organized in a core and in recommended modules. The *Core Ontology* consists of the hierarchy of upper and general types, i.e. those that meet a large consensus across languages and provide the most essential information for describing word senses, whereas the so-called *Recommended Ontology* includes the hierarchy lower and specific types that clearly provide more granular information about word meaning. Language/application-specific semantic types may also be designed in order to allow for a more refined description level.

We will not enter in the description of each ontological nodes of the SIMPLE-CLIPS ontology and the interested reader is referred to (Ruimy *et al*., 2003) for a detailed description. We think that it is sufficient for our aim to provide a figure representing the Ontology (Fig. 3).

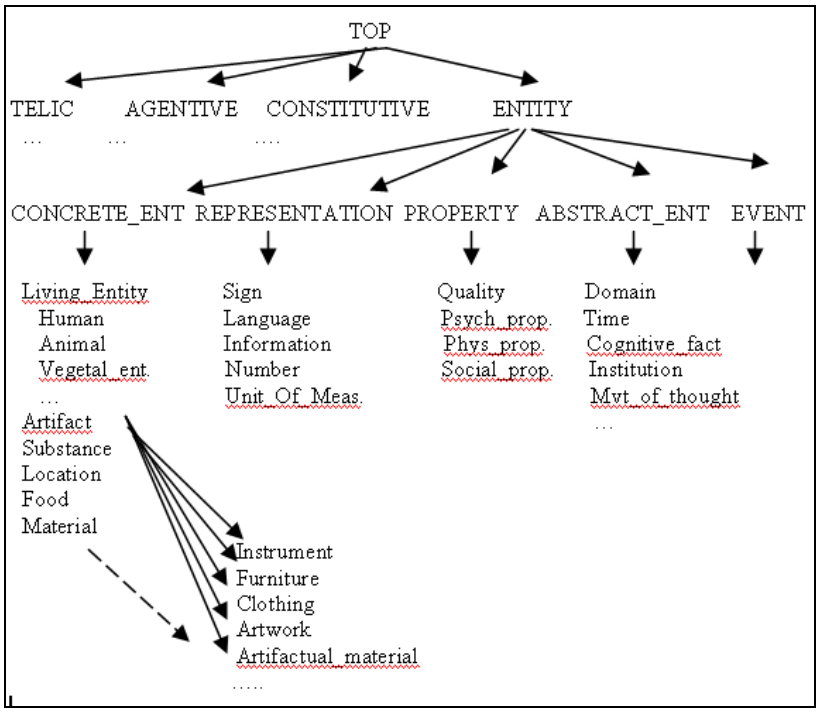


Fig. 3: the SIMPLE-CLIPS Ontology

1.1.2.2 TEMPLATES IN SIMPLE-CLIPS

In SIMPLE, the encoding process was guided by the use of the so-called *templates*, i.e. schematic structures allowing a semantic type to be constrained to a structured cluster of information considered crucial to its definition. We consider the use of Templates to be very interesting: in fact, their use should ensure encoding uniformity and consistency, thus providing a way to create resources well-suited for NLP.

1.1.2.3 THE ITALIAN SIMPLE-CLIPS LEXICON

The Italian SIMPLE-CLIPS Lexicon consists of semantic entries of verbs, nouns and adjectives, described using a wide variaty of information types:

- Type assignment and type hierarchy information
- Domain information
- Qualia Structure
- Encoding of regular polysemy
- Synonymy
- Derivational information
- Semantic Features
- Argument structure (and linking to syntactic entries)

In the following, we will briefly describe these information types and try to render the level of complexity and the richness of this lexicon. For a more detailed description of the expressive modalities in SIMPLE-CLIPS and also for a discussion of the problems that emerged during the encoding phase, the interested reader is referred to (Ruimy *et al*., 2003).

1.1.2.3.1 Type assignment

Assigning a semantic type to a lexicon entry implies the inheritance of the type hierarchy information. This information indicates the position of the type, and hence of the SemU which instantiates such a type, within the whole type hierarchy. It is provided by means of a feature whose attribute depends on whether the semantic type the entry belongs to is a SIMPLE-CLIPS one or a unified one (for which the inheritance path, which consists of both a supertype and a telic or/and agentive dimension of meanings, is made explicit).

1.1.2.3.2 Domain

Domain classes (a set of 350 domains) supply information on the topic of texts in which the SemU at hand is more likely to occur.

1.1.2.3.3 Qualia Structure

Formal

The formal role provides a broad characterization of an entity with respect to other entities, expressed by the 'isa' hyperonymic relation for SIMPLE-CLIPS nouns and event-denoting entities. The 'isa' relation supplies a more granular type of information than the semantic type and allows a further subtyping of entries sharing the same template, like for example: 'isa' rettile, felino, pachiderma (reptile, feline, pachyderm) enable to differentiate and subclassify entries encoded in the EARTH_ANIMAL type. As a general rule, the closest hyperonym is thus assigned and circular 'isa' relations avoided as far as possible.

For adjectives, following WordNet and in contrast to the encoding of nouns and verbs, the formal role is not expressed by hyperonymy but rather by an antonymic relation.

Constitutive

The constitutive role expresses the internal constitution of an entity by means of a set of relation of the type: 'is_a_member_of', 'is_a_part_of', 'has_as_member', 'resulting_state', 'has_as_property'. A special mention should be made for the constitutive relation 'concerns' that is largely used. In the template type DISEASE, for example, this relation is used, whenever possible, to indicate the organ affected by the disease e.g.: for congiuntivite (conjunctivitis) → occhio (eye). Similarly, some semantic units typed as CLOTHING are assigned as target of the 'concerns' relation uomo (man) or donna (woman). Another use of this relation is the specification, in the entry of words denoting shops, of the main item offered for sale: libreria → libro (bookshop → book).

The constitutive quale plays a particularly crucial role in the semantic description of adjectives. It is the place where meaning components, which are essential to capture the adjectival meaning, are expressed in terms of features (for example 'movement', 'space', and 'substance' etc.).

Agentive

The agentive role provides information on the origin of an entity. Typical agentive relations that were used in the Italian lexicon are 'created_by' (for all kinds of artifacts), 'result_of'; 'caused_by', 'agentive_prog, 'agentive_cause', 'agentive_experience'.

Telic

The telic role specifies the function of an entity, the purpose for which it exists or has been created. The main telic relations instantiated are the following ones: 'used_as', 'used_for', 'is_the_activity_of', 'object_of_the_activity', 'indirect_telic', 'telic'. Agentive and telic roles are never instantiated in the adjective description since they are considered to be expressing semantic dimensions of the noun rather than of the adjective.

We think that this stronger "relational" part of the database is the one that most will serve the purposes of the QA application. For this reason, we will also provide in the next tables an overview of all the available "semantic relations" for the four qualia roles (Table 2, Table 3, Table 4, Table 5).

Name	Description	Example	Isa
Formal	Formal node in the hierarchy		Top
Isa	<SemU2> is the hyperonym of <SemU1>. The value of this relation can be given, for example, by a EuroWordNet hyperonym or by a dictionary superordinate;	*Isa* (<yacht>, <boat>)	*Formal*
Antonym_comp	<SemU2> is the complementary antonym of <SemU1>	*AntonymComp* (<dead>, <alive>)	*Formal*
Antonym_grad	<SemU2> is the gradable antonym of <SemU1>	*AntonymGrad* (<hot>, <cold>)	*Formal*
Antonym_mult	<SemU2> is one of the multiple antonyms of <SemU1>	*AntonymMult* (<German>, <Dutch>)	*Formal*

Table 2: SIMPLE-CLIPS relations for the Formal role

Name	Description	Example	Isa
Constitutive	Formal node in the hierarchy		Top
Is_a_member_of	<SemU1> is a member or element of <SemU2>. <SemU1> is typically a shaped, countable entity, and <SemU2> is typically a collective entity, i.e. a set of individuals	*Is_a_member_of* (<senator>, <senate>)	*Constitutive*
Has_as_member	<SemU1>, which corresponds to a collective entity or a set of entities, has <SemU2> as its (proto)-typical member or element	*Has_as_member* (<flock>, <bird>)	*Constitutive*
Is_a_part_of	<SemU1> is a part of <SemU2>	*Is_a_part_of* (<head>,	*Constitutive*

		<body>)	
Has_as_part	<SemU1> has prototypically <SemU2> as one of its parts	*Has_as_part* (<airplane>, <wing>)	*Constitutive*
Location	Formal Node in the hierarchy		*Constitutive*
Property	Formal node in the hierarchy		*Constitutive*
Instrument	<SemU1> is an event SemU and <SemU2> is the typical instrument, vehicle or device which is used to perform this event.	*Instrument* (<ski>, <ski>)	*Constitutive*
Relates	<SemU1> denotes a relation, and <SemU2> denotes the typical entities that are related by it	*Relates* (<kinship>, <person>)	*Constitutive*
Resulting_state	<SemU1> is a transition and <SemU2> is the resulting state of the transition	*Resulting_state* (<die>, <dead>)	*Constitutive*
Is_a_follower_of	<SemU1> is an individual who is a follower, a supporter, an adept of a certain religion, doctrine, school of thought or credo in <SemU2>	*Is_a_follower_of* (<marxist>, <marxism>)	*Is_a_member*
Made_of	<SemU2>is typically a substance or stuff out of which <SemU1> is made. Alternatively, <SemU2> is an element which enters into the composition of <SemU1>	*Made_of* (<bread>, <flour>); *Made_of* (<water>, <oxigen>)	*Is_a_part_of*
Is_in	<SemU1> is typically located in <SemU2>.	*Is_in* (<oasis>, <desert>)	*Location*
Lives_in	<SemU1> is a living entity which typically lives in <SemU2>.	*Lives_in* (<Italian> <Italy>)	*Location*
Has_as_colour	<SemU2> is the typical colour of <SemU1>	*Has_as_colour* (<lemon>, <yellow>)	*Property*
Constitutive_activity	<SemU2> is the typical activity of <SemU1>, which is a natural kind entity and the subject of the event expressed by <SemU2>	*Constitutive_activity* (<bird>, <fly>)	*Property*
Produces	<SemU2> is a natural entity that is typically produced by <SemU1>, which is also a natural kind entity	*Produces* (<bird>, <egg>)	*Property*
Produced_by	<SemU1> is an entity that is typically produced by <SemU2> as the result of a natural process, intrinsically correlated with the nature of <SemU2>.	*Produced_by* (<honey>, <bee>)	*Property*
Property_of	<SemU2> is an adjective which refers to the property, quality or attribute expressed by <SemU1>	*Property_of* (<intelligence>, <intelligent>)	*Property*
Concerns	<SemU1> is a phenomenon, event or situation that typically concerns of affects <SemU2>	*Concerns* (<hepatitis>, <liver>)	*Property*
Contains	<SemU2> is an object which is typically contained in <SemU1>	*Contains* (<book, information>)	*Property*
Quantifies	<SemU1> expresses a quantity of <SemU2>	*Quantifies* (<bottle>, <liquid>)	*Property*
Measured_by	<SemU1> is a property which is measured by <SemU2>, a unit of mesure	*Measured_by* (<temperature>, <degree>)	*Property*
Related_to	<SemU1> is related in some unspecified way to <SemU2>	*Related_to* (<second>, <two>)	*Property*
Successor_of	<SemU1> is the element following <Sem2> in a series	*Successor_of* (<two>, <one>)	*Property*
Has_as_effect	<SemU2> is a side-effect, consequence or indirect effect of <SemU1>	*Has_as_effect* (<storm>, <thunder>	*Property*
Typical_of	<SemU1> is a disease or phenomenon that typically affects the entity in <SemU2>	*Typical_of* (<distemper>, <dog>)	*Property*
Causes	<SemU1> typically causes <SemU2> as part of its natural constitution	*Causes* (<measles>, <fever>)	*Property*

Table 3: SIMPLE-CLIPS relations for the Consititutive role

Name	Description	Example	Isal
Telic	Formal node in the hierarchy	*Telic* (<pet>, <company>)	Top
Direct_telic	Formal node in the hierarchy		*Telic*
Indirect_telic	<SemU1> and <SemU2> are related through an underspecified indirect telic relation. <SemU1> is usually the subject or the instrument-complement of the event in <SemU2>, which represents a purpose prototypically associated with <SemU1>	*Indirect_telic* (<eye>, <see>)	*Telic*
Purpose	<SemU1> is the SemU being defined, and <SemU2> is an event corresponding to the intended purpose of <SemU1>	*Purpose* (<send>, <receive>)	*Telic*
Object_of_the_activity	<SemU2> is an event whose direct object is typically <SemU1>, and expresses an activity which is the typical purpose of <SemU1>.	*Object_of_the_activity* (<book>, <read>)	*Direct_telic*
Activity	Formal node in the hierarchy		*Indirect_telic*
Instrumental	Formal node in the hierarchy		*Indirect_telic*
Is_the_activity_of	<SemU2> is the characterizing activity of <SemU1>	*Is_the_activity_of* (<doctor> <heal>)	*Activity*
Is_the_ability_of	<SemU2> is a typical ability of an individual in <SemU1>	*Is_the_ability_of* (<painter>, <paint>)	*Activity*
Is_the_habit_of	<SemU2> is the typical habit of an individual in <SemU1>	*Is_the_habit_of* (<smoker>, <smoke>)	*Activity*
Used_for	): <SemU2> is the typical function of <SemU1>. This relation usually applies to instruments or devices to connect them with the activity in which they are used or to their typical purpose.	*Used_for* (<crane>, <lift>)	*Instrumental*
Used_by	<SemU1> is typically used by <SemU2>	*Used_by* (<lancet>, <surgeon>)	*Instrumental*
Used_against	<SemU1> is used typically against <SemU2>	*Used_against* (<chemiotherapy>, <cancer>)	*Instrumental*
Used_as	<SemU1> is typically used with the function which is expressed by <SemU2>	*Used_as* (<wood>, <material>)	*Instrumental*

Table 4: SIMPLE-CLIPS relations for the Telic role

Name	Description	Example	Is_a_l
Agentive	Formal node in the hierarchy	*Agentive* (<student>, <study>)	Top
Result_of	<SemU1> is an entity which is the result, effect or by-product of the event expressed by <SemU2>	*Result_of* (<loss>, <loose>)	*Agentive*
Agentive_prog	<SemU2> is an event which is ongoing while an individual has the property expressed by <SemU1>	*Agentive_prog* (<pedestrian>, <walk>)	*Agentive*
Artifactual_agentive	Formal node in the hierarchy		*Agentive*
Agentive_Cause	<SemU1> is a causative verb, and <SemU2> is the causing component of the event	*Agentive_Cause* (<sink>, <cause>)	*Agentive*
Agentive_Experience	<SemU1> is an experience predicate and <SemU2> is the event experienced by the individual.	*Agentive_Experience* (<fear>, <feel>)	*Agentive*
Caused_by	<SemU1> is a phenomenon or natural event which is produced by <SemU2>	*Caused_by* (<infection>, <bacterion>	*Agentive*
Source	<SemU2> is the source or origin of <SemU1>	*Source* (<law>, <society>)	*Agentive*
Created_by	<SemU1> is obtained, or created by a certain human process or action <SemU2>	*Created_by* (<book>, <write>)	*Artifactual_agentive*
Derived_from	<SemU1> is derived from another object <SemU2>	*Derived_from*	*Artifactual_age*

	through a certain process of alteration	(<petrol>, <oil>)	*ntive*

Table 5: SIMPLE-CLIPS relations for the Agentive role

1.1.2.3.4 Regular Polysemy

In the SIMPLE-CLIPS lexicon, senses of nouns that are systematically related are described according to a set of 20 well-established sense alternation classes. For verbs, the phenomenon of regular polysemy concerns the inchoative and causative alternation of predicates. As for adjectives, two regular polysemous classes are identified, i.e. the alternation between 'nationality' and 'style' and between 'temperature' and 'behaviour'. This kind of sense ambiguity, which is referred to as logical polysemy in Pustejovsky's theory (Pustejovsky, 1995) and gives rise to Complex types, has been captured and represented in SIMPLE-CLIPS by linking the different SemUs of a lexical item entering a regular polysemous class. The link is expressed in each relevant template by the name of the pair of semantic types to which the alternative senses belong.

1.1.2.3.5 Synonymy

A synonymic relation is assigned to those SemUs encoded in top templates and for which taxonomic information (expressed by a formal relation) does not make sense like, for example, parte (part), scopo (goal), mezzo (means), maniera (manner), etc. Synonymic relations are also used in the encoding of adjectives - especially for highly polysemic ones.

1.1.2.3.6 Derivational information

Cross-categorial links such as derivation are marked by means of relations linking the derived semantic unit to the base one. A set of relations allows differentiating between various types of derivation, namely deverbal and denominal adjectives, location or instrument denoting deverbal nouns, deadjectival nouns, denominal nouns, nouns derived from proper nouns, deverbal nouns, denominal verbs, verbs derived from instrument-denoting nouns, subject, object and indirect object nominalizations of verbs, nominalizations of verbs. The following table sums up the available derivational relations.

Name	Description	Example
Nounadjective	<SemU1> is a noun which derives from the adjective in <SemU2>: *The <SemU1> of X => X is <SemU2>*	*Nounadjective* (<intelligence>, <intelligent>)
Agentverb	<SemU1> is an agentive noun, which lexicalizes the agent argument of the verb in <SemU2>	*Agentverb* (<writer>, <to write>
Patientverb	<SemU1> is a noun which lexicalizes the patient argument of the verb in <SemU2>	*Patientverb* (<employee>, <to employ>)
Eventverb	<SemU1> is an event nominal, and refers to the event expressed by the verb in <SemU2>	*Eventverb* (<destruction>, <to destroy>)
Stateverb	<SemU1> is a noun which refers to a state which either	*Stateverb* (<hate> <to

	is expressed by the verb in <SemU2>, or is the result of the event expressed by the verb in <SemU2>	hate>
DenominalVerbNoun	<SemU1> is a noun from which the verb in <SemU2> derives	*DenominalVerbNoun* (<to butter>, <butter>)
Processverb	<SemU1> is a process nominal, and refers to the process expressed by the verb in <SemU2>	*Processverb* (<thought>, <to think>)
NounPropernoun	<SemU2> is a proper noun from which <SemU1> derives	*NounPropernoun* (<Marxism>, <Marx>)
NounNoun	<SemU> is a noun deriving from another noun <SemU2>	*NounNoun* (<Communist>, <Communism>)

Table 6: SIMPLE-CLIPS derivational relations

1.1.2.3.7 Semantic Features

The use of some semantic features enables the retrieval and clustering of entries encoded in different semantic types but still sharing a common meaning component, e.g.: 'plus_collective', 'plus_edible' etc.

1.1.2.3.8 Argument structure

One of the most interesting aspects in SIMPLE-CLIPS is the possibility of encode lexical predicates for each predicative SemU (verb, deverbal, deadjectival or SIMPLE-CLIPS noun). For verbs and SIMPLE-CLIPS predicative nouns, predicate names coincide with the SemU naming. As to deverbal nouns, they share with their verbal base the same predicate, i.e. accusare, *accusatore*, *accusato*, *accusa* (to accuse, accuser, accused, accusation) all point to the predicate *accusare*.. Each argument is assigned a semantic role (selected in a predefined list of roles based on EAGLES recommendations) and the information concerning its semantic characterization. This information is clearly not to be taken as a real restriction but rather as a preference of combination, in prototypical situations.

1.1.3 Complementary and overlapping information types in IWN and SIMPLE-CLIPS

The two lexicons under analysis have been constructed at different times and within the framework of different European and national projects. Their linguistic design is very different: ItalWordNet is a semantic net whose building block is the synset, while the basic unit in the SIMPLE-CLIPS lexicon is the Semantic Unit, i.e. the word sense. This difference was also noted in a preliminary comparison of the two resources presented in (Roventini *et al*., 2002). In the next table (Table 7), we report the list of lexical-semantic information types that can be found in computational lexicons (and are also recommended in the EAGLES guidelines for lexical semantics encoding), with which we opened this chapter. We will indicate what information and features are present in IWN and SIMPLE-CLIPS, in order to represent the level of complementary and overlapping between the two resources.

Information Type	Description	ItalWordNet	SIMPLE-CLIPS
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	Example of use in context	✓	✓

Table 7: lexico-semantic information types in IWN and SIMPLE-CLIPS lexicons

The table shows that the two lexicons seem more or less to share the same information types. The most important difference seems to be the lack of any type of argument structure representation in IWN and of multilinguality in SIMPLE-CLIPS. These distinctions are not completely valid: as a matter of fact, IWN incorporates in its design the possibility of encoding ROLE/INVOLVEMENT relations that can be used to represent something similar to semantic roles (agent, patient, location etc.) even if an object "semantic frame" is completely missing. Moreover, it is true that multilingual information is not present at the moment in the SIMPLE-CLIPS database, but its design is completely open to the possibility of add a further multilingual layer to the already encoded morphological, syntactic and semantic ones (and we have to remember that the SIMPLE-CLIPS design has provided the model of the MILE, the Multilingual ISLE lexical entry, Calzolari *et al.*, 2002).

The truth is that the differences are there but they are fuzzier than the ones that can be captured by means of a SIMPLE-CLIPS yes/no table. As far as the IWN database is concerned, the choice of adopting the semantic net and the synset as representational devices determined some important consequences: first of all, the synsets are supposed to represent concepts and not separate word senses. This means that the members of the same synset share the same hyperonym, the same Top Concept, the same definition and also all the possible relations with other synsets in the net. In SIMPLE-CLIPS, on the contrary, we see that the choice of representing meaning in the form of a SemU determined that less attention was paid in the coherent taxonomical construction of the lexicon. Synonymy has not been encoded in a pervasive way but only

"assigned to those SemUs [..] for which taxonomic information [..]does not make sense" (Ruimy *et al.*, 2003). This means that often two equivalent (synonym) SemUs don't have anything that connects them, not even the same hyperonym. It may be a problem for the automatic exploitation of the SIMPLE-CLIPS information in QA applications, since the possibility of navigating the SemUs by following hierarchical lines is something very useful for QA, as much as the exploitation of alternative forms of the same concept (that we should be able to derive from synonyms). If these are the "positive" features of IWN with respect to SIMPLE-CLIPS, however, it has to be said that SIMPLE-CLIPS is surely a more complete and rich lexicon, covering an impressive amount of lexicon-semantic information types, for example a very interesting connection with a syntactic layer of representation (which IWN completely devoid).

Moreover, SIMPLE-CLIPS may be able to overcome the weakness of its taxonomical structures by recurring to the almost 160 different Semantic Types which constitute its rich and detailed Top Ontology (versus the only 60 Top Concepts of the EWN/IWN Top Ontology) and which may be an important answer in supporting inferences requiring generalizations.

We also have to remember that a certain difference in size exists between IWN and SIMPLE-CLIPS (respectively 65,000 versus 57,000 overall senses) and that this difference can have an impact on the final application.

In the following chapter, we will introduce Question Answering and we will see which types of lexico-semantic information are used in the most advanced systems.

2

In this chapter, we will introduce the topics that play the principal role in this dissertation: Question Answering and lexico-semantic language resources. After a brief "historical account" of the QA, we will introduce the features that constitute the backbone of a QA application. We will concentrate our attention in particular on the modules of the overall architecture that, more than others, constitute environments where lexico-semantic information can be exploited.

2.1 Open-Domain 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 and the CLEF campaigns.

The Text REtrieval Conference[12] (TREC) is a series of workshops organized by the National Institute of Standards and Technology (NIST), devised to continue improving state-of-the-art Information Retrieval (IR). Within TREC, a series of evaluations of English Question-Answering systems has started in 1999 (TREC-9).

The Cross-Language Evaluation Forum[13] (CLEF) on the other hand consists in annual evaluation campaigns and workshops with the aim of stimulating the development of mono- and multilingual information retrieval systems for European languages and of contributing to the building of a research community in the multidisciplinary area of multilingual information access. Since 2003 CLEF has also been offering the mono and multilingual QA track among a series of tracks designed to test different aspects of mono- and cross-language system performance.

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.

[12] trec.nist.gov
[13] www.clef-campaign.org

In the following we will introduce different examples of current systems and "historical" applications, but first of all we would like to highlight the numerous different dimensions that determine the complexity of QA. We will go along (Hirschman and Gaizauskas, 2001), who explicitly 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.

2.2 The many dimensions of the Question Answering problem

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* Hirschman and Gaizauskas (2001) 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.

Another useful way to realize how many different issues complicate the research on QA is to consider the document which presents the first QA Roadmap (Burger *et al*., 2001). In that document a "deliberately ambitious vision for research in Q&A" is outlined, in order "to define the program structures capable of addressing the question processing and answer extraction subtasks and combine them in increasingly sophisticated ways".

The vision for research in QA is graphically presented in the Roadmap recurring to the following figure (Fig. 4):

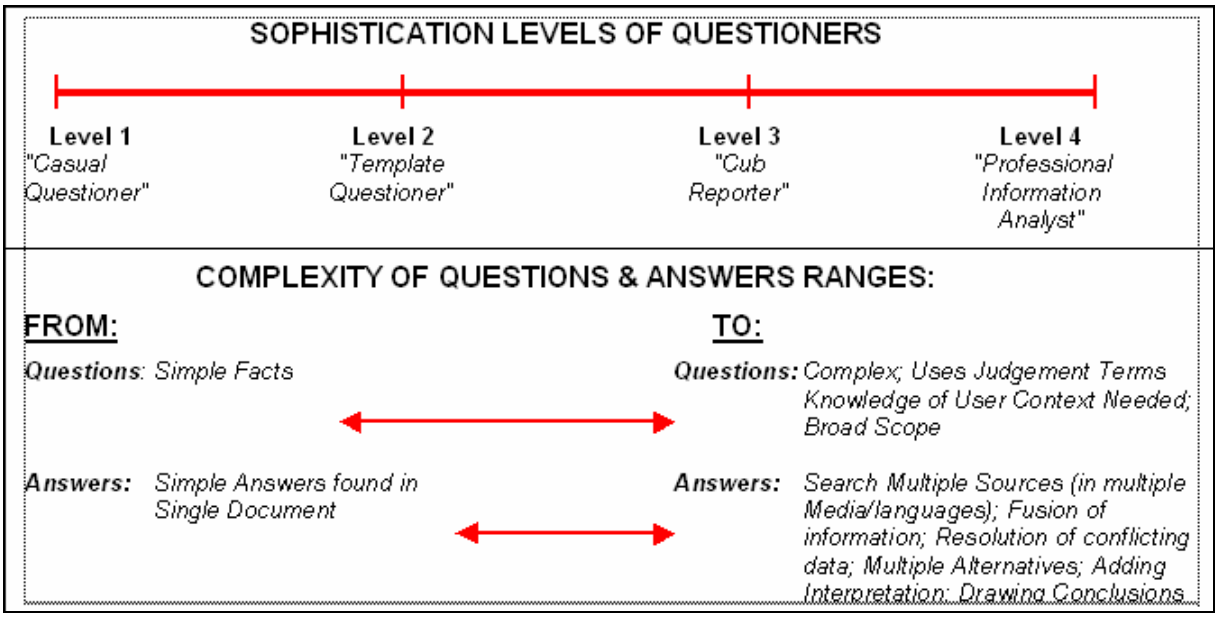


Fig. 4: spectrum of questioner, question and answer types delimiting complexity of the QA problem (Burger *et al.*, 2001)

It is possible to see how three main axes are identified that determine the level of complexity for QA applications: on one hand, the "user profile", which goes from the simplest "Causal Questioner" (i.e. the common, non-professional user of the Internet that needs to individuate answers to general question) to the more skilled "Professional Information Analyst", which is an expert of a domain and needs something able to provide non-trivial answers resulting also from reasoning and treatment of implicitness. The other axis is the type of question, that ranges from the simple factual question (the one that is now proposed to the TREC and CLEF participants as test for their systems) to the complex questions which require the treatment of opinions expressed by the questioner, the possibility of dealing with the pragmatic context of the questioner and also of covering broader scopes in the presentation of the question focus. The last axis is the one pertaining to the answer types that range form the simple answer found in a single document to the answers extracted from multi-media and multilingual collections, or answers that mean dealing with multiple alternatives, interpretation, and summarization of the textual content.

In the first Roadmap for QA, twelve lines of research are identified that in some way intersect the "dimensions of the problems" listed in (Hirschman and Gaizauskas, 2001); along these lines of research it becomes necessary to study different types of questions (question taxonomies) and the various aspects connected to question processing (understanding, ambiguities, implicatures and reformulations), the role of the context, the data sources for QA, the aspects connected to answer extraction, the importance of defining the user profile for QA etc.

Moreover, during an important workshop (Maybury *et al.*, 2002) held within the LREC-2002 conference, a second Roadmap was defined. In the resulting document (Maybury, 2002) the set of dimensions that distinguish various question answering systems are presented considering that systems typologies range from application for on-line help to access encyclopaedias or technical manual, to open web-based question answering, to sophisticated QA in support of business or military intelligence analyses.

In the second Roadmap an extensive (even if not yet complete) list of characteristics that distinguish QA environments is presented:

- the nature of the query, including the question form (e.g., keyword(s),phrase(s),full question(s)) the question type (e.g., who, what, when, where, how, why, what-if), and the intention of the question e.g., request, command, inform).
- the level of complexity of the question and answer,
- characteristics of the source(s)and/or supporting corpora (e.g., size, dynamicity, quality),
- properties of the domain and/or task (e.g., degree of structure, complexity),
- the potential for answer reuse,
- the degree of performance required (e.g., precision and recall),
- the nature of the users (e.g., age, expertise, language proficiency ,degree of motivation)and the importance of usability,
- the purposes of the users (e.g., help with homework or cooking, strategic analysis),
- nature of supporting knowledge sources (e.g., degree of necessary linguistic ,world knowledge)
- reasoning requirements (e.g., inference required for question analysis, answer retrieval, presentation generation)
- the degree of multilinguality and crosslinguality (e.g., questions might be asked in one language),
- the user model (e.g., stereotypical vs. individualized user models)
- the task model (e.g., structured vs. unstructured tasks)
- the type of answers provided (e.g., named entities, phrases, factoid, link to document, summary)
- the nature of interaction (e.g., user reactivity, mixed initiative, question and answer refinement, answer justification)

All these dimensions and features concur to create the suggestive figure (Fig. 5) which graphically represents the many issues and problems paving the Roadmap from 2002 until 2006. The three main *lanes* deal with: i) resources necessary to develop or evaluate QA systems, ii) methods and algorithms, iii) and systems. A common, long term outcome of the roadmap are high quality QA systems. Each lane then leads to outcomes such as measurable progress from having shared resources, a QA toolkit, and personalized QA. Intermediate results are typology of users, topology of answers, a model of QA tasks, QA reuse across sessions, and interactive dialogue. Roadblocks along the way include the need to manage *user expectations*, the need for reusable test collections and evaluation methods. Since the difficulty of natural language processing (NLP) and inference has limited the scope of QA, these were represented as speed limit signs. On the right hand side of the road map we can see the progression of question (from simple *factoid* questions to *what-if* questions) and answer types (from simple facts to multimodal answers). The Roadmap also indicated related fields (requiring cross-community fertilization) such as high performance knowledge bases (HPKB) (Cohen *et al*, 1998), topic detection and tracking (TDT), databases, virtual reference desks, and user modelling that

were indicated by the workshop participants as having particular importance for solving the general QA problem.

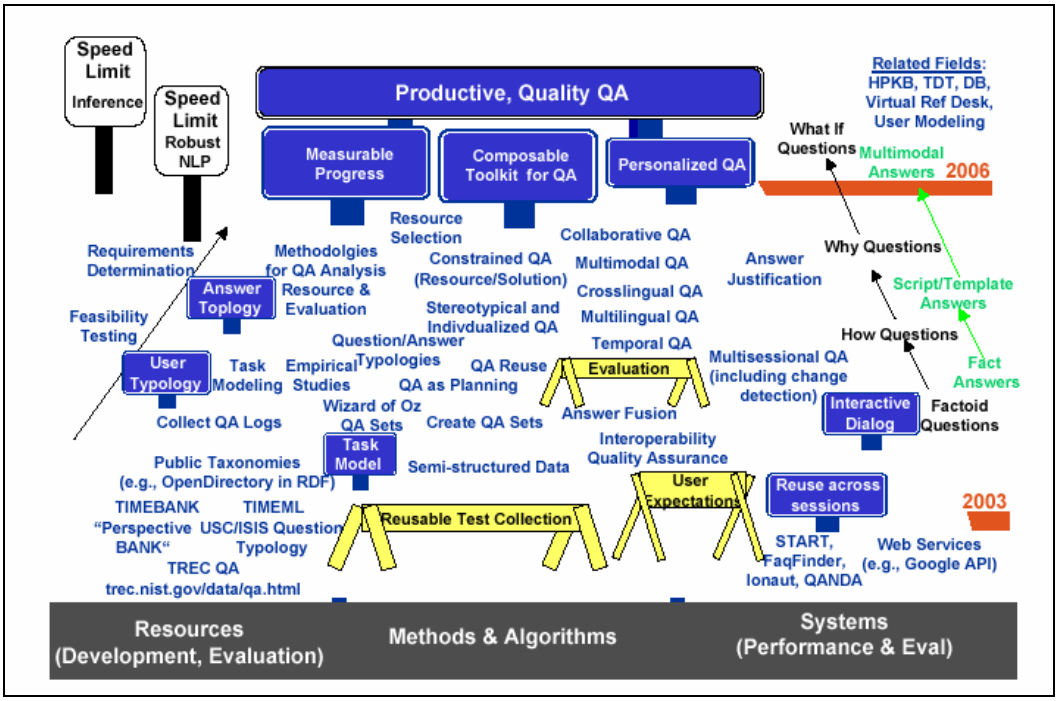


Fig. 5: Roadmap jointly created by participants of the LREC 2002 Q&A workshop

All this dimensions cut across the QA problem, determining a very large variety of possible instantiation of the same "application type". We will not take into account all these dimensions, on the contrary, given the aim and the scope of this dissertation, we will circumscribe a very small field of action. We thought it was important, however, to provide an exemplification of how difficult it is to talk about QA as it would be a unique focus of study and on the contrary how different the possible solutions, approaches and final results are that one can achieve.

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.

2.3 History and types of QA

In order to provide a brief historical account of QA, we will refer to the surveys presented in (Hirschman and Gaizauskas, 2001) and in (Monz, 2003b). (Hirschman and Gaizauskas, 2001) also reports on

the work by (Simmons, 1965) that, by the middle of the 60s, already illustrated about fifteen implemented English language question-answering systems built over the previous five years.. Simmons (1965) defined Question Answering as a term rather loosely used…

> *"to include general-purpose language processors which deal with natural English statements and/or questions. These vary from conversation machines to machine which generate sentences in response to pictures, and systems which translate from English into logical calculi. All of these may be interpreted in some sense as attempting to use natural English in a manner very closely related to the question and answer pattern"*

(Simmons, 1965) already recognized different sub-types of QA applications: i) the List-Structured Data-Base Systems (which deal with data organized in list form), ii) Graphic Data-Base Systems (based on graphic databases) and, most important for the future development of the field, iii) the Text-Based Systems, i.e. systems that attempt to find answers in ordinary English text: PROTOSYNTHEX (Simmons and McConlogue, 1963), ALA (Thorne, 1962) and General Inquirer (Stone *et al*., 1962) was elected by (Simmons, 1965) as representative of this class.

Also 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).

Given the aim of this dissertation and its stress on knowledge sources we will choose the knowledge-source type as the discriminating factor which helps us to classify the different approaches to QA. For this reason, we will consider together the conversational agents and the front-end systems surveyed by (Hirschman and Gaizauskas, 2001) because both approaches employ the same knowledge source type (i.e. structured information stored in a database of facts).

2.3.1 Front-Ends to Structured Knowledge Repositories

These systems were (and still are) intended as an 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, 1977) but also PHLIQA1 (Bronnenberg *et al*., 1980), surveyed by (Monz, 2003b). They were not toy systems (LUNAR was demonstrated at a science convention in 1971 and was able to answer more than 90% of the questions) but were intended for very restricted domains: BASEBALL answered questions about games of the baseball American league, STUDENT was designed to solve algebra problems, LUNAR answered questions about moon rocks and soil samples gathered during the Apollo 11 lunar mission, and PHLIQA1 was designed to answer short questions about computer installations in Europe.

We want to mention here other two systems which exploited information stored in database: SHRDLU (Winograd, 1972) and GUS. These two systems are particular because they are devised as dialogue interactive advisory systems.

Aim of the SHRDLU and GUS systems was helping researchers to study and understand issues connected to human dialogue (such as the treatment of anaphora and ellipsis). SHRDLU simulated a robot moving objects while GUS simulated a travel advisor; we prefer to consider these two systems in this category even if (Hirschman and Gaizauskas, 2001) states that the typology of source of knowledge (in this case structured knowledge and not free text) was not a distinctive and necessary feature of these systems, that had as its main goal the definition of strategies for human-computer interaction for which real-time response was a fundamental requirement. Tradition of research dedicated to conversational agents has flowed in the current line dedicated to spoken language interfaces, whose exemplar result is the Jupiter system by MIT, which provides a telephone-based conversational interface for weather information (Zue *et al*., 2000).

In general, the design, architecture and modules of QA conceived as natural language front-end to Structured Knowledge Repositories are very interesting; usually in these systems different steps of analysis of the question are foreseen, ranging from the morphological to the syntactic and semantic analysis. These systems analyse the question and, exploiting linguistic knowledge, transform it into a canonical form used to generate a formal query that is matched against the database content.

But 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.

Therefore, we will introduce a distinction between two different types of QA[14] that during the years have specialized and differentiated their employed techniques and strategies:

- 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.

The next category of QA systems is dedicated to the approaches that, over the years, have confronted the challenging task of extracting precise information from open-ended collections of unstructured texts. This approach is also the one that most paves the way for modern Open-Domain QA systems.

[14] For a comparison of the different requirements of the two QA types, cf. (Rinaldi et al., 2003).

2.3.2 Text-Based Question Answering

In this approach to QA, the source of knowledge is not a database where facts are described in particular formats and formalisms but free text in plain-text format[15]. The text could be a huge collection of documents of various nature (newspaper articles, encyclopaedias, collection of news and facts available on line, etc.) or a single text. Textual question answering systems work by matching textual units in the question with textual units in documents. The interesting work by C. Monz (Monz, 2003b) provides a detailed analysis of many systems that can be ascribed to this category, such as ORACLE (Phillips, 1960), PROTOSYNTHEX (Simmons *et al.*, 1963), Wendlandt and Driscoll's system and MURAX (Kupiec, 1993) and the interested reader is referred to his surveys for a more in-depth discussion. Here, still referring to the Monz's survey, we will briefly introduce the features of these systems that we think are most interesting with respect to the scope of this dissertation. For this reason, we will mention the modules exploiting linguistic and especially semantic information.

The first system presented in the Monz's survey is ORACLE, which produces a syntactic analysis of both question and the text where the answer can be contained. The question is transformed into a declarative sentence and the new word order is sought in the corpus. The semantics had a very small role in this system, where only few entities were labelled with tags indicating time or places.

PROTOSYNTHEX is interesting not only because the textual material where the answer is searched is an encyclopaedia, but also because the corpus is syntactically analysed using a dependency parser, an approach which has survived and very frequently used in current QA systems (Attardi *et al.*, 2001, Harabagiu *et al.*, 2000, but also the system developed within this research, cf. chapter 4.

Also the Automatic Language Analyzer (ALA) (Thorne, 1962) exploits a formalism that is strongly related to the dependency graphs. But what is most interesting in this system is that there is an attempt to exploit lexical knowledge by taking into account a measure of *semantic correlation* that, as Monz suggests, has a strong resemblance to the notion of mutual information in (Fano, 1961).

Thematic roles based on the work by Fillmore are instead exploited in the Wendlandt and Driscoll's system (Wendlandt and Driscoll, 1991): the system tries to recognize thematic roles (such as *agent*, *object*, *instrument* etc.) and attributes (abstract categories such as *amount*, *size*, *order* etc.) which occur in the question and in a ranked list of paragraphs closest to the question. The thematic roles and the attributes are used to compute a similarity score based on the common roles and attributes.

We also want to count in this same group a system which is classified by (Monz, 2003b) in a separate category dedicated to the so-called Inference-Based System: the Semantic Information Retriever (SIR) (Raphael, 1964). In SIR, the input text is transformed into a kind of logical form that can be queried by the user. What can be important from our point of view is that this system tries to exploit a first group of lexical semantic relations, such as PART-OF and IS-A, which are instantiated, for example, transforming the

[15] Current systems deal also with more structured type of text data, i.e. in HTML or XML format, but in this case a pre-processing phase is envisaged: the textual collection is indexed and only textual material are returned to the system.

input text *every boy is a person* in the relation SETR(boy, person), which means that *boy* is a subset of *person*.

In the survey by (Hirschman and Gaizauskas, 2001) we see that the reference system for this type of approach to QA is W. Lehnert's QUALM (Lehnert, 1977), that is instead counted in the Inference-based systems class in Monz's survey. Lehnert's work can be ascribed to the studies on *human story comprehension* and, even if it is almost thirty years old, it still constitutes a fundamental reference for many researchers working in the QA field, in particular because it was the first to provide an extensive treatment of question classification, a methodology that is now fully entered in the design of current QA systems.

Lehnert defined a complete theoretical framework for QA, where either the question or the answer text are subjected to the same type of analysis, aimed at building conceptual dependency representations that are matched onto each other in order to provide answer to the question. But the matching between question and answer at conceptual dependency level does not exhaust the entire process, that is instead driven and constrained by the recognition of the type of question. Lehnert determined thirteen question categories, of which we provide a brief explanation given the importance the "question classification" issue will have in our work (and also for a comparison we will do between these categories and categories most used in modern QA). For Lehnert, questions can be classified in the following classes :

- Causal antecedent: questions asking about states or events that have in some way caused the concept in question (examples are Why did John go to New York? Or How did the glass break?)
- Goal Orientation: special cases of the so-called Why questions, it includes questions asking about motives and goals behind an action (examples are *Mary left for what reason? For what reason did John take the book?*)
- Enablement: specify a causal relationship of the type ENABLE between the question concept and an unknown act or state that enables it (example are *How was John able to eat?* And *What did John need to do in order to leave?*)
- Causal Consequent: causal structures in which the question concept causes an unknown concept or causal chain. The relation expressing this link is called LEADTO. Examples are *What resulted from John's leaving? What happened when John left?*
- Verification: questions asking about the truth of an event. Lehnert says that they more or less correspond to yes/no questions and are represented as single concepts with a MODE value. Examples are *Did John leave?* And *Does John think that Mary left?*.
- Disjunctive: the same as Verification questions but foreseeing multiple concepts instead of single. Examples are *Was John or Mary here?*
- Instrumental/Procedural: questions asking about the procedure or instrument connected to the question concept. Examples are *How did John go to New York?* and *What did John use to eat?*

- Concept Completion: they include many Who, What, Where and When questions and are a kind of fill-in-the-blank question because they ask about the completion of an event with a missing knowledge component. Examples are *What did John eat?* and *Who gave Mary the book?*
- Expectational: ask about the causal antecedent of an act that presumably did not occur, therefore they often have the form of Why-Not questions. Examples are *Why didn't John go to New York? Why isn't John eating?*
- Judgemental: questions that solicit a judgment on the part of the listener (example are: *what should John do to keep Mary from leaving?* and *What should John do now?*).
- Quantification: question asking about an amount of something (examples are *How many people are here?* but also *How ill was John?*)
- Feature Specification: questions asking about properties of a given person or thing. Examples are *What colour are John's eyes?* And *What breed of dog is Rover?.*
- Request: this type of question is not used when (as happens with all the other question types) a person wants to obtain specific information but rather when someone wants someone else to perform some act. Example are *Would you pass me the salt?* and *Can you get me my coat?*.

All these question types are discerned by a question analyser that functions as a discrimination net which applies distinction rules on conceptual graphs. It has to be said that categories are not mutually exclusive since obviously in many cases the same question can be classified according to more than one class. Lehnert recognizes that the most difficult category to individuate is the Question Completion, that is characterized by an unknown conceptual component that can occur everywhere at any level of conceptualisation. To understand how difficult it is to recognize this type of question, we have to realize that the recognizing strategy consists in this case in selecting a Question Completion type only after all the other possibilities have been discarded.

Another characteristic that makes W. Lehnert's work a kind of "classic" for those practicing QA, is the study of how inferential analysis can provide other constraints on answer selection. Inferential analysis examines the content of a question to see whether the initial categorization is correct. In order to accomplish this difficult task, Lehnert makes pragmatics and world-knowledge play an important role. So, in the case of the question *Do you have a wooden match?*, the ultimate goal of the QA system is not to infer that the expected answer is *yes* or *no* (as the general rule on Verification questions would suggest) but rather to understand that the questioner want a match to light something (attributing in this way the question to the Request category). Some rule-based inference mechanisms are thus individuated with the aim of achieving conceptual understanding of the question.

The attempt to deal with not literal meanings of the question, the definition of question type classes, the very complexity and breadth of the entire theoretical framework which is aimed at providing either a psychological and computational model for QA; all these things make of Lehnert's work something unparalleled in the survey of literature on QA.

Inspired by Lehnert's results, other researchers in the community of *story comprehension* (the field that studies the way humans understand stories and are able to answer questions about them) have defined alternative approaches to QA. In this field we already mention the work of A. Graesser and his research group at the University of Memphis (Graesser and Murachver, 1985)[16].

The story comprehension work and the current QA on text-collection systems both have in common the characteristic that answers must be derived from unstructured texts. But, as (Hirschman and Gaizauskas, 2001) point out, unlike text-collection QA, the text containing the answer is known in advance. Moreover, multiple questions about a single text force and allow text processing at a deeper level than what it is possible to perform and achieve when the system has to deal with massive numbers of texts. At the same time, the story comprehension environment seems to provide less answer redundancy than QA on open text collection and this aspect could increase the difficulty of the answer location task.

For systems working on free-text, the most important issue is the possibility of pinpoint, among the many textual units, the one which more probably constitutes the answer to the question. In order to accomplish this task, what many text-based systems have in common is an information retrieval system that extracts document(s), or part of documents, containing the answer(s) to the question. In this sense the IR module serves the purpose of individuating a subset of documents or paragraphs that became *candidate answers* and that will be further analysed to more precisely locate the actual answer[17].

But IR is just one of the many modules constituting the overall architecture of modern Open-Domain QA systems. In the next paragraph, we will introduce a generic architecture for automatic QA.

2.4 A Generic Architecture for QA

In order to outline a general skeleton for QA, we will refer to the overviews proposed in (Hirschman and Gaizauskas, 2001; Paşca, 2003; Monz, 2003b). The systems that every year are presented to the TREC and CLEF QA tracks are very diverse for typology of design and techniques. Anyway, researchers generally agree that QA architecture in constituted by distinct modules for question processing, passage retrieval and answer processing (Abney *et al.*, 2000; Hovy *et al.*, 2001 and Moldovan *et al.*, 2000). The following modules are conceived to belong to a QA application that accepts natural language questions as input and has, as a knowledge source, a large collection of natural language texts. The question is analysed and a set of candidate answers is extracted from the text collection. As a last step, the answer is identified among the candidates. All these unique aspects constitute the lowest common denominator for most current QA systems. Thus, the fundamental modules for a coarse and general QA architecture are[18]:

[16] Graesser and his collaborators added several new categories to the set identified in QUALM, among others the classes "comparison", "definition", "interpretation" and "example" that can be useful when coming to modern QA systems.

[17] This step is very important for the whole process of textual QA and it will be further discussed in the following chapter.

[18] A Document Collection Pre-processing is an indispensable but implicit phase and is not included in the generic logic architecture.

- Question Analysis
- IR module
- Answer Selection and Extraction

The first step consists in the (morphosyntactic and/or dependency-based and/or semantic) analysis of the natural language question posed by the user. The question is usually classified according to an ontology of question types with the aim of determining what the expected answer will be (for example a place, a date, a human name etc).

Part of the output of the question analysis step (the query obtained by the question) is the input for the next phase, generally coinciding with an Information Retrieval module, which identifies documents (or paragraphs) that contain terms of the query (the so-called *candidate answers*). Other information, such as the question category and syntactic roles, are instead sent to the last module. The retrieval component returns a set of (usually ranked) documents/paragraphs that will be further analysed in the last step.

Answer Selection and Extraction module takes the set of candidate answers as input, together with additional information which resulted from the question analysis phase. At the end of this last processing phase, a phrase is selected that is most likely to be a correct answer and returned to the user.

Fig. 6 provides a schematic representation of this information and processing flow. More details will be provided in the following chapter, when we introduce the QA prototype built within the current research.

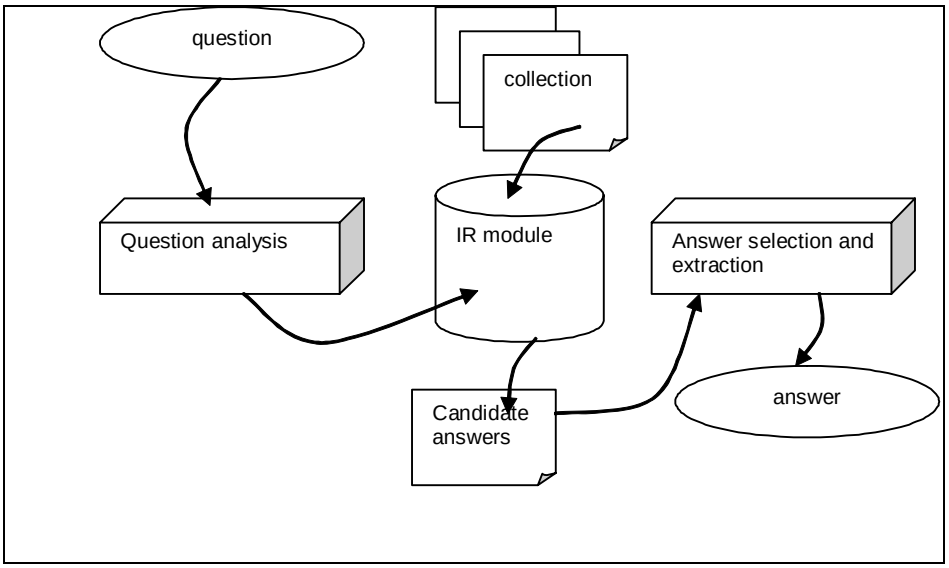


Fig. 6: a generic QA architecture

2.5 Language Resources contribution to QA systems

Given the aim of this dissertation, we think it would be useful at this point to provide a preliminary overview of the exploitation of language resources in various QA systems. In order to prepare this overview we will refer to more than twenty systems participating in the (English) QA track of the 12th edition of the TREC

conference (TREC-2003). We will see that, even if we generically talk about the involvement of "language resources"in QA, WordNet is basically the only lexicon exploited in most of the surveyed systems[19]. Moreover, WordNet exploitation seems somehow limited to the use of its hyperonyms (in the question classification phase) and its synonyms (basically in the query expansion module). Only the last three systems (Massot *et al*., 2003; Paranjpe *et al*., 2003; Harabagiu *et al*., 2003) seem capable of exploiting other type of relations and information available in WordNet.

2.5.1 TREC-12 systems with lexico-semantic feedback

In the system of the University of Sheffield (Gaizauskas *et al*., 2003) WordNet plays an important role in judging the correctness of the answer and in performing query expansion. WordNet is used to compute the proximity of the candidate answer with the entity sought by the question (the expected answer type) (following the idea that the closer the two items are in the semantic net, the more the candidate answer is likely to be the sought answer). Furthermore, WordNet provides the sets of terms that can be used to perform query expansion in the passage retrieval phase. Non only are synonyms of the question terms considered but so are the terms that are in the WordNet glosses.

In the ISI system, TextMap (Echihabi *et al*., 2003), the knowledge-based module exploits different information sources and also the WordNet lexicon, whose hierarchical links are used to support the answer selection and the strategies to handle the DEFINITION question. Moreover, the sources of knowledge used by the knowledge-based answer selection module proved to have a stronger impact on the overall performance of the answer selection system than the ability to automatically train parameters in the pattern- and statistics-based systems, which use poorer representations.

WordNet is also used to validate correctness in definitional questions in the system by BBN Technologies (Xu *et al*., 2003) that verifies that the question type is a hypernym of the answer. In that system WordNet is also exploited to match verbs (for example, "Who killed X?" = "Y shot X") even if the adopted methodology is not clear.

The system developed at the National University of Singapore (Yang *et al*., 2003), QUALIFIER, performs Event Mining to discover and then incorporate the knowledge of event structure (describing different facets of the event, like time, location, object, action etc.) for more effective QA. The semantic gap between the query space and document space is filled with knowledge of lexical resources to expand the original query. The new query therefore contains terms that are related to the lexical context in WordNet.

The Carnegie Mellon system, Javelin (Nyberg *et al*., 2003), uses hypernym and meronym relationships in WordNet to determine the link between candidate answers and the target answer type.

Many other systems make use of WordNet information, in general in the question classification phase and in the query expansion by means of synonyms. This happens for example in the systems of

[19] Only the system developed at the IBM T.J. Watson Research Center (Prager et al., 2003) exploits also the CYC ontology.

University of Amsterdam (Jijkoun *et al*, 2003), of the ILS Institute (Wu *et al*., 2003) and in the DIOGENE system (Kouylekov *et al*., 2003).

WordNet synonyms are also exploited for query expansion by the systems of the University of Wales (Clifton, 2003), Queens College (Grunfeld, 2003) and of the Massachussets Institute of Technology (Katz *et al*., 2003). In this last system, moreover, the list of *occupations* from WordNet (such as *actor*, *spokesman*, *leader* etc.) was used to boost the precision of the module which recognizes pattern of the type occupation+human name.

Two very interesting approaches exploiting all the information available in the form of semantic relations in WordNet are represented by the system of the University of Catalunya (Massot *et al*., 2003) and the one of the Indian Institute of Technology of Mumbai (Paranjpe *et al*., 2003).

The system of the Universitat Politècnica de Catalunya makes use of the list of synsets (with no attempt to Word Sense Disambiguation), the list of hyperonyms of each synset (up to the top of each hyperonym chain), the EWN's Top Concept Ontology (Rodriguez *et al*,1998), the Domain Code (Magnini and Cavaglià, 2000) and a list of relations actor-action obtained through an analysis of the glosses of WordNet. This information is learned by an automatic classifier and used to pinpoint the correct answer (with poor results).

The system described in (Paranjpe *et al*., 2003) is instead aimed at building *Bayesian network* on all the WordNet lexical relations able to represent inference.

The Language Computer Corporation QA system (Harabagiu *et al*., 2003), in order to improve precision, makes use of a theorem prover that produces abductive justification of the answer by accessing a axiomatic transformation of the WordNet glosses. WordNet is also used to classify the type of question and determine the type of the expected answer. This last system is the one which performed best and it appears to be the most knowledge-intensive. The functioning of its modules that most exploit lexico-semantic information will be extensively described in this chapter.

2.5.2 The interface between QA and language resources

We would like to propose the generic architecture schema once again (Fig. 7). This time we add three callouts with the purpose of zooming in on the portions of the general architecture where semantic information is usually more exploited. These callouts are important because they represent the “articulated joint” between the two logical components of this dissertation: the QA application on one side, and lexico-semantic language resources on the other.

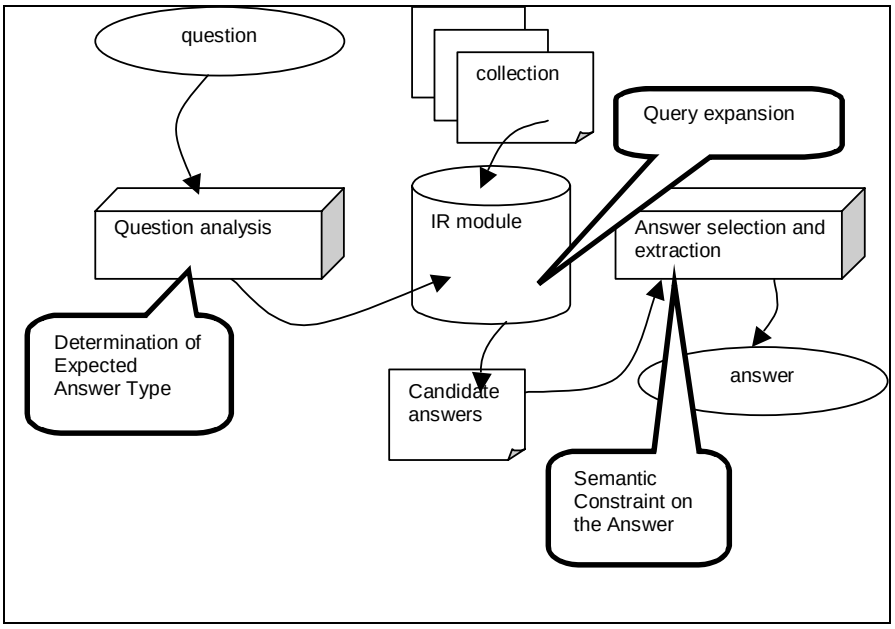


Fig. 7: LRs exploitation on a generic QA schema

The reference work for the exploitation of lexico-semantic language resources in QA applications is surely represented by the research carried out by Sanda Harabagiu, Dan Moldovan and their collaborators at the University of Texas in Dallas and within the activity of a private company (Language Computer Corporation, LCC). The system developed by this group, FALCON, is one of the systems that participated in the TREC-12 and the one that obtained the best results. In what follows, we will provide a detailed description of the methodologies and techniques for the exploitation of semantic information adopted by this research group. What we think is important is that in the work of this group the awareness of the importance of the lexico-semantic feedback in natural language applications is clearly evident. (Harabagiu *et al*., 2001) reports the results of the TREC-9 evaluation (Kwok *et al*., 2000 and Radev *et al*., 2000) as evidence of the fact that Information Retrieval techniques alone are not sufficient to find precise answers. If questions are treated simply as vectors of words, so following a consolidated technique in Information Retrieval (the vector-space model, cf. Lee *et al*., 1997), the overall performance of the QA system is quite low (Berger *et al*., 2000; Clarke *et al*., 2000). Thus many systems that participated in recent years to the TREC and CLEF conferences adopted architectures which attempted to capture the semantics of a question in order to exploit it in answer extraction.

QA has inspired new research in the challenging integration of surface-text-based methods with knowledge-base text inference. A QA system needs to capture the semantics of open-domain questions and also to justify the correctness of answers. In order to make *semantics* play a role in the entire QA pipeline, information in language resources can be successfully mined, as it is shown in (Paşca and Harabagiu, 2001).

2.5.2.1 LEXICO-SEMANTIC FEEDBACK IN QUESTION ANALYSIS

When processing a question, the main goal is being able to recognize the question type and the *expected answer type* (i.e. what the question is looking for). Many existing systems demonstrated the usefulness of

recognizing the expected answer type in answer extraction (Abney *et al*., 2000; Srihari and Li., 2000; Kouylekov *et al*., 2003; Attardi *et al*., 2001 *inter alia*). This information can often be derived by the so-called *question stem* (the *Wh*-element, i.e. the interrogative adverb, adjective or pronoun at the beginning of the sentence). In fact, there is generally a strong correlation between the *question stem* and the *expected answer type*. For example, if the question stem is *Who*, we would expect the answer to be the name of a person while if the question stem is *When* we can expect to find a temporal expression etc.

The following table (extracted from Paşca, 2003) shows an exemplar distribution of queries on the basis of the question stems of the TREC question test collection:

QUESTION STEM	PCT.	SAMPLE QUESTION
What	48%	*What is the life expectancy of an elephant?*
Who	18%	*Who was the architect of the Central Park?*
Where	10%	*Where is Romania located?*
How	9%	*How hot is the core of the Earth?*
When	8%	*When was the Triangle Shirtwaist fire?*
Name	3%	*Name a film in which Jude Low acted?*
Which	2%	*Which U.S. President is buried in Washington, D.C?*
Why	1%	*Why can't ostriches fly?*
Whom	1%	*Whom did the Chicago Bulls beat in the 1993 championship?*

Table 8: Distribution of question stems for the TREC test collection (from Paşca, 2003)

But the derivation of the semantic category of the expected answer cannot be carried out solely on the basis of question stem. In fact, semantically equivalent questions may be introduced by different stems (***Who*** *interprets Mulder in the X-Files?* Vs ***What*** *actor interprets Mulder in the X-Files?*) while, at the same time, the same stem may ask about completely different categories (***What*** *was the country that invaded Poland in 1939?* Vs ***What*** *mammal lives in the Ocean?*).

The problem is that sometimes the question stem is ambiguous: a certain level of ambiguity is for example present in the stem *Who*, since it can introduce questions asking about the name of a person (*Who is the president of the USA?*) but also about a group (*Who produced the Panda?*) or about the role of someone (*Who is Silvio Berlusconi?*).

The more ambiguous the question stem is, the more difficult it is to process questions in such a way that its representation drives answer extraction. (Voorhers, 1999) demonstrated a correlation between the lower precision scores and the level of ambiguity introduced by a question stem. The most ambiguous stems are *What* and *Which,* that can be used to introduce questions that can have many types of expected answer (such as locations, humans, weights, abstract entities etc.).

In those cases, what disambiguates the question is the so-called *answer type term* (ATT), i.e. the term preceded by the ambiguous question stem that allows the derivation of the expected answer type. The answer type term is usually a noun (but it can also be a verb or an adjective) and is extracted by recurring to

syntax-based rules. In the following table, again extracted from (Paşca, 2003), some questions of the TREC test collection are accompanied by their answer type term and their corresponding expected answer type (we also provide the Italian translation).

QUESTION	QUESTION STEM	ANSWER TYPE TERM	EXPECTED ANSWER TYPE
What was the name of the Titanic's captain? *Qual era il nome del capiano del Titanic?*	What Quale	Captain Capitano	Human
What U.S. Government agency registers trademarks? *Quale agenzia del Governo Americano registra i marchi di fabbrica?*	What Quale	Agency Agenzia	Organization
What is the capital of Kosovo? *Qual è la capitale del Kosovo?*	What Quale	Capital Capitale	Town
What state does Charles Robb represent? *Quale stato è rappresentato da Charles Robb?*	What Quale	State Stato	Province
How much does one ton of cement cost? *Quanto costa una tonnellata di cemento?*	How Much Quanto	Cost costare	Money
How long did the Charles Manson murder trial last? *Quanto durò il processo per l'assassinio di Charles Manson?*	How long Quanto	Last durare	Quantity
What is the population of Japan? *Qual è la popolazione del Giappone?*	What Quale	Population Popolazione	Number

Table 9: TREC questions represented through question stems and expected answer types (Paşca, 2003)

If the QA system has access to extensive, open-domain lexico-semantic resources, the recognition of the expected answer type is feasible for a broad range of fact-seeking questions (Paşca and Harabagiu, 2001). In cases like these in fact, the semantic category of the expected answers is derived by projecting the question dependency representation onto an answer type hierarchy encoding lexico-semantic information available in language resources. Here we want to report the solutions adopted by Harabagiu and her collaborators within the FALCON QA system, described in many papers (cf. for example Harabagiu *et al*., 2000, Paşca and Harabagiu, 2001) and presented in detail in (Paşca, 2003). This same methodology has been adopted in many other current QA systems (for example Attardi *et al*., 2001 and Magnini *et al*., 2001) and relies on a hierarchy of answer types that incorporates morphological, lexical and semantic information available in the WordNet database (Miller, 1990).

2.5.2.1.1 The FALCON Hierarchy of Answer Types

The hierarchy designed for the FALCON (Harabagiu *et al*., 2000) system is three-pronged: the first level is formed by the semantic category nodes corresponding to the answer types. The second level consists of links

from answer types to WordNet sub-hierarchies. The mappings of each hierarchy leaf node onto a named entity category represent the lowest part of the hierarchy (Paşca, 2003).

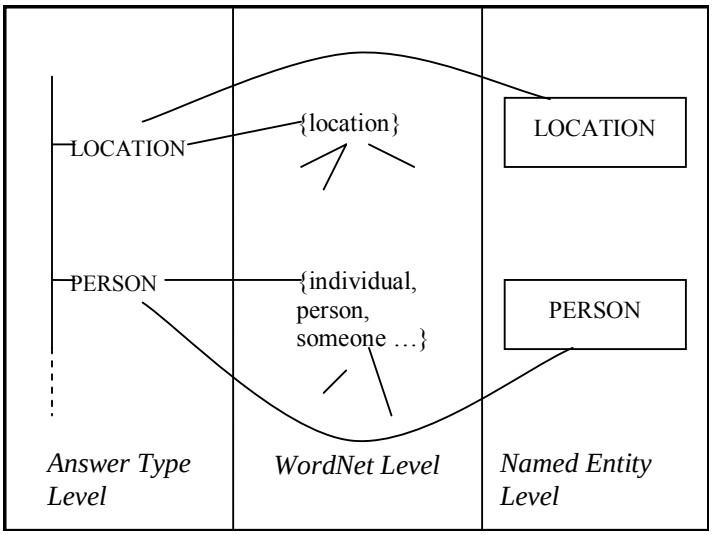


Fig. 8: articulation of the FALCON Hierarchy of Answer Types

The first two layers of the hierarchy are motivated by the fact that the semantic taxonomies in WordNet are not structured according to the categories frequently arising in questions, so the general-purpose lexical hierarchies have to be re-interpreted in the light of the QA requirements.

(Paşca, 2003) introduces the details which prevent the WordNet classes from being chosen as they are as nodes of the answer type hierarchy:

- semantically related entities are occasionally not grouped under the same category in WordNet (so for example *Mount Etna* and *Mount Elbert* are grouped under different hyperonyms, namely *location* and *object*).
- WordNet semantic categories are too general to give a useful categorization of entities for QA (and that's true for classes such as *feeling*, *object*, *artefact* etc.)

The answer type terms of the upper level of the hierarchy give a sense of generality while the lower level embeds information from WordNet under the appropriate categories. In the following figure, extracted from (Paşca and Harabagiu, 2001), we can see how different the nodes are of the Expected Answer Type Taxonomy from the top nodes in WordNet 1.6.

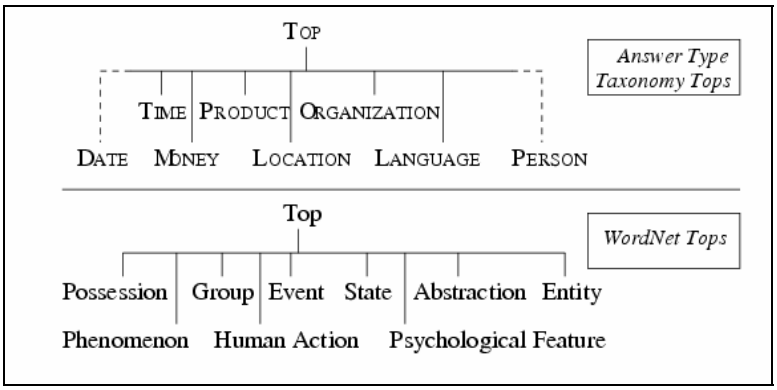


Fig. 9: Example of Answer Types nodes and WordNet top nodes (from Paşca and Harabagiu, 2001)

The nodes in the upper level are chosen in such a way to be general enough to ensure wide coverage of the various possible questions and are independent from any particular domains. The selection of these nodes is based on two sources of information: WordNet semantic domains and named entity recognition.

The "population" of the nodes of the Expected Answer Type taxonomy is described in details in (Paşca, 2003). The general idea is that the IS-A relation can be successfully exploited in the attempt to avoid to manually add all the answer type terms under the highest nodes of the Expected Answer Type hierarchy. The IS-A allows us to recognize common denominators among the answer type terms, in such a way that, for example, in the case of the questions:

i) *What French oceanographer owned the "Calypso"?*,
ii) *What biologist founded the science of genetics?*
iii) *What scientist discover the vaccine against Hepatitis-B?*

the word *scientist* can be used to gather *oceanographer* and *biologist* under the same Expected Answer Type. The underlying assumption is that all the WordNet sub-hierarchies having *scientist* as a root are also going to define a PERSON's name. Fortunately, at least for what the top PERSON is concerned, all the WordNet nouns referring to "types of human being" have a common generalized concept and this feature allowed Harabagiu's research group to exploit the taxonomic links by connecting only this common concept to the Top of the Expected Answer Type hierarchy[20].

Following this same strategy, all the nodes in the hierarchy have been populated with WordNet sub-hierarchies. Human intervention was needed in the construction of the WordNet-based hierarchy. Repeated experiments with increasing sets of questions suggested how to enlarge and refine the hierarchy. The following figure (Fig. 10) shows all the WordNet sub-hierarchies that are gathered under the same Expected

20 In (Paşca, 2003) we find an exemplar picture which partially shows the WordNet sub-hierarchies linked to the node person. In that picture, we see that the sub-hierarchies are led by the synsets {scientist, man of science}, {European}, {philosopher}, {inhabitant, denizen, dweller}, {guardian, defender}, {performer}, etc. In the following chapter, where we introduce the similar experience we had in building the so-called Answer Type Taxonomy for Italian questions, we will demonstrate that the fine-grained distinctions that characterize the taxonomy used by the FALCON system in sometimes not required. It's not clear, in fact, why the WordNet synset {person, individual, someone, somebody, mortal, human, soul} has not been directly linked to the person node, preferring instead to select more fine-grained sub-taxonomies, that could inherit the general node person from their intermediate hyperonyms.

answer Type MONEY. In this figure, it is possible to see how different parts of speech may specialize the same answer type since verbs and nouns equally contribute to the successful recognition of the answer type of questions like *How much could you rent a Volkswagen bug for in 1966?* and *What is the monetary value of the Noble Prize in 1989?*.

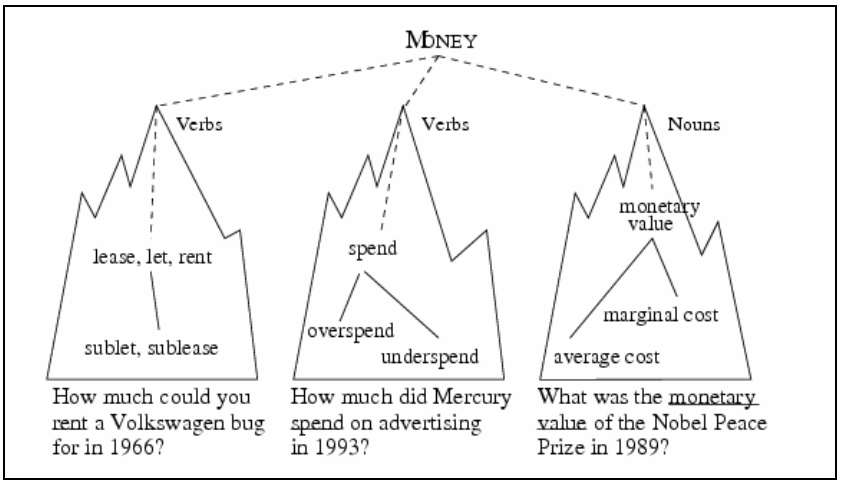


Fig. 10: WordNet sub-hierarchies collected under the Expected Answer Type MONEY (from Paşca and Harabagiu, 2001)

But not only can nouns and verbs be ATT, so can adjectives, in particular in conjunction with the question stem *How* (*how many*, *how wide*, *how long*, *how tall* etc.). In that case, FALCON exploits a WordNet relation different from the IS-A, i.e. the ATTRIBUTE/VALUE_OF relation, which links adjectives with their corresponding property (so *tall* is linked with *stature*, *long* with *length* etc.). In this way the relation allows the system to transform the adjective into the corresponding noun synset that is used to access the hierarchy.

Another important feature of the answer type hierarchy is that it partially deals with the important issue concerning the *word sense disambiguation* (WSD) of the answer type term. In fact, the ATT can be a polysemous word, thus "distributed" in different WordNet synsets. In this way, the specific requirements of the QA and the scope of factual questions, the sub-hierarchies often gather only some of the senses of the word or even several words with different senses (by linking them under a unique node of the hierarchy).

Another thing that must be noted is that there is a many-to-many relation between the named entity category and the leaves of the answer type top hierarchies. In the example reported in (Harabagiu *et al*., 2000) we see that the answer type node MONEY is searched either as the *money* or as the *price* named entity category. In contrast, the named entity category quantity is used to recognize four types of answers, i.e. SPEED, DURATION, DIMENSION and AMOUNT.

The next figure, on the other hand, shows instead an important and pervasive characteristic of the architecture of the expected answer hierarchy (and in a way its very *reason of being*), i.e. the fact that the top layer is just used to collect and gather diverse taxonomical portions that can reside in scattered parts of

WordNet. The structure allows the system to treat in the same way both questions containing the ATTs *wingspan* and *size* (i.e. looking in the text for a named entity of the type "dimension"), even if they reside in different part of the semantic net.

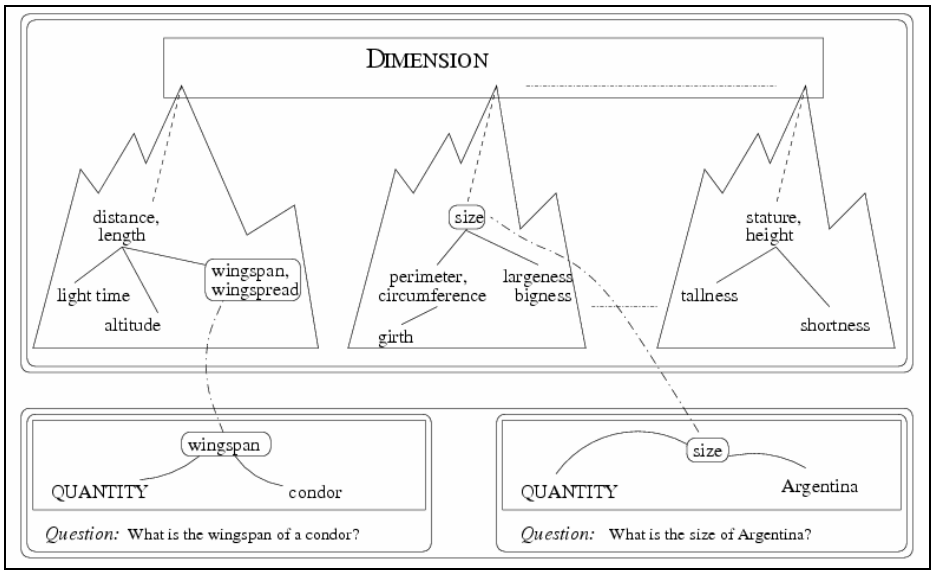


Fig. 11: mapping of the dimension leaf in several WordNet classes (from Harabagiu *et al.*, 2000)

One of the "big issues" of QA is surely the one concerning the possibility, once that the expected answer has been recognized, to distinguish in the text particular words that can be traced back to their expected answer class. It means that, given the question *What nation hosted the Olympic Games in 2004?*, even if the system has been able to "understand" that the expected answer is the name of a country, this is not a useful information unless the system is also able to detect the entities of the type "country" in the text. So, the answer type hierarchy used in the FALCON system also foresees a third component consisting in named entity categories supported by a given *Named Entities Recognizer* (NERec). FALCON is supported by a NERec which recognizes sixteen NE categories (for example *date*, *product*, *human*, *province*, *organization*, *country*, *time*, *city* etc.), that have been mapped onto the leaf nodes of the expected answer hierarchy. This means, for example, that the node CITY is mapped into a corresponding named entity category CITY thus allowing the recognition of city instances such as Los Angeles, Hamburg etc. (cf. always Paşca, 2003). The highly modular architecture of the expected answer hierarchy and the logical separation between the layer concerning lexical nodes and named entity categories makes the hierarchy independent from the underlying recognition implementation. At the same time, in case a better recognition technology becomes available, it can be integrated changing only the interface between the two layers without having to change the entire hierarchy.

2.5.2.1.2 Deriving the type of expected answer in FALCON

In (Paşca, 2003) a detailed description of the way the expected answer type is derived in FALCON by exploiting the hierarchy is provided. The ATT is extracted by using a dependency syntactic representation of the question and by following the *principle of syntactic proximity*, that states that usually the *question stem* and *the answer type term* are situated in relative proximity to each other in the dependency representation. In fact, for the majority of questions, the answer type term and the question stem are directly connected to each other through a relation which places them in immediate syntactic proximity.

Exceptions are the semantically redundant terms such as *name* in *What is the name…?* or *type* in *Which type of …?*. These terms can be safely ignored and treated as special cases of stop terms, that are useful only because they serve, in the dependency representation, as intermediate connectors between the semantic salient ATT and the question stem (this is the case, for example, of the question *What is the name of the highest mountain in Africa?* for which the system is able to recognize *mountain* as ATT).

Once the ATT has been successfully identified, the expected answer type hierarchy nodes are iteratively inspected. (Paşca, 2003) uses the example question *What French oceanographer owned the "Calypso"?*. The ATT *oceanographer* is searched on all the WordNet sub-hierarchies. In fact *What*, differently from other question stems such as *Who* or *When*, cannot be used to select any specific expected answer type. *Oceanographer* is founded as a hyponym of the noun synset {scientist, man of science} which is the root of one of the WN sub-hierarchies embedded in the node PERSON. Finally, the expected answer type is stored in the question stem node of the dependency representation, representing in this way the disambiguation of the previously ambiguous question stem with the semantic category of the possible answers.

2.5.2.1.3 Dynamic Answer Type Categories in FALCON

(Paşca, 2003) describes the situation in which the expected answer type definition strategies may fail, i.e. when the semantic type does not correspond to a specific named entity but rather to a common noun[21]. For example, in the question *What flower did Vincent Van Gogh paint?*, the NERec would not be able to detect *sunflowers* in the text (as usually happens), thus condemning this type of question to have an unknown answer type.

The strategy adopted in FALCON is one of the most successful examples of exploitation of WordNet in the NLP field: answer types are populated with synsets collected from their WordNet hyponyms and, if any of the thus collected hyponyms are in relevant document fragments, it becomes a candidate answer.

So, for the exemplar question *What flower did Vincent Van Gogh paint?*, the system generates a dynamic answer type "flower" populated with 470 WordNet hyponyms (e.g. *sunflower*, *petunia*, *orchid* etc.),

[21] In (Paşca, 2003) this situation is described as: " when the submitted questions asks about semantic categories that are too specific to be captured in a separated named entity category". We do not really think this it the real difference, since the problem is not the specificity of the noun but rather the fact that instance of the ATT is likely to be found in the text not as a named entity (a temporal expression, a Proper Name etc.) but rather as a common noun.

identifying in this way the text fragment *In March 1987, van Gogh's "Sunflowers" sold for $39.9 million at Christie's in London*.

The system is also provided with a mechanism to decide whether to use an existing "static" answer type or a "dynamic", "on-the-fly" one. First of all, if using the normal procedure the system is able to derive an expected answer type linked to a named entity category, then that answer type is used in the process. If the question contains cue words indicating specialization (such as *type*, *kind*, *sort*, *variety* etc.), then the expected answer type is created on the fly.

2.5.2.2 KNOWLEDGE-BOOSTED PASSAGE RETRIEVAL

The second module where lexico-semantic language resource can be exploited is the Information Retrieval component that is represented as the core of the entire general architecture of Fig. 3. Information Retrieval is not only the core of the QA architecture because it is somehow in the middle between the question and the answer processing modules, but also because its effectiveness is strategically important for the overall performance of a question answering system. In fact, if the document retrieval component does not return any document containing an answer, even perfectly functioning question analysis and answer extraction modules will obviously fail to return a correct answer to the user. An IR system works a representation of the document. The chosen representation provides therefore what can be regarded as a particular semantic interpretation of the document. The task of information retrieval consists in matching the semantic interpretation of the document with the one expressed by the user in the query. Almost all existing IR systems simply represent documents and queries as a 'bag-of-words'. This is not adequate to the most challenging tasks, but it is effective for simple retrieval tasks.

Generally speaking, the aim of the information retrieval component is not to find specific answers to the question, but to identify documents that are likely to contain an answer. It's a kind of pre-selection of documents also known as *pre-fetching* (Monz, 2003b).

Current QA systems often employ i) a boolean retrieval component, which provides more options in query formulation, and ii) paragraph/passage-based retrieval that seems more suitable in QA because the answers are normally expressed very locally in a document and also because short text excerpts are easier to process by later components of the question answering system. The impact of passage-based retrieval vs. full document retrieval was positively demonstrated by (Llopis *et al.*, 2002), while (Montz, 2003) arrived to different conclusions.

The basic idea is that the question "enters" the IR module in the form of a list of keywords and a set of candidate answers with a paragraph-long text span are returned by the IR module.

It is quite obvious that the selection of the "right" keyword to send to the IR module is crucial in order to obtain a useful set of candidate answers. In (Paşca, 2003) we find an inventory of the factors that drive the selection of question terms as keywords. These factors are: semantic salience, redundancy and degree of term variation. Usually, however, given the difficulty in automatically assessing these aspects, the unifying criterion used to identify the keywords to submit to the IR module is only the morphosyntactic

information expressed by the Part of Speech. In the following chapter, where we describe the QA Italian prototype built within the current research, we also propose a scalable semantic criterion that can be used in the selection of the keyword with some success (see 4.3.3.3).

Most current QA systems exploit lexical-semantic information when dealing with term variation. Term variation is indeed one of the major difficulties for QA and in general for every application having the aim to access natural language texts. Terms may vary morphologically (when we find the verbal form *go* in the question and *went* in the answer), lexically (when in the question we find *car* and in the answer we find *automobile*) or semantically (when the *oil-tanker* of the question is different from the more general *ship* in the answer).

An experiment carried out by Moldovan and Harabagiu's research group (Paşca and Harabagiu, 2001) shows that WordNet can play a very important role in the keyword selection and expansion phases. First of all, it seems that a correlation between *semantic salience* and *specificity* of the keyword can be established and that very specific keywords should not be dropped from the query.

WordNet can provide information about the *specificity* of the keyword, specificity that is assessed by off-line counting the hyponyms of the concept.

Furthermore, WordNet has an important informative role in generating keyword variation by exploiting synonyms.

(Paşca and Harabagiu, 2001) provides the results of the quantitative and qualitative evaluation carried out on the TREC-9 QA test collection (consisting of 3 Gb of documents) by using 893 questions constituting the TREC-8 and TREC-9 test sets; as far as the "specificity issue" is concerned, the experiment shows that, when the specificity option is enabled, the number of the TREC-8 correctly answered questions increases from 133 to 151. Also the results of the experiment carried out to evaluate effectiveness of the strategies dealing with keyword variation show a significant improve: the precision was 55.3% if no alternation was enabled, 67.6% if the lexical alternation was allowed and 73.7% if both lexical and semantic alternation was enabled. What the paper does not say is how the problem of word sense disambiguation is handled in these experiments, since choosing the "wrong sense" prior to performing the query expansion has showed (Sanderson, 2000) to have very a negative impact on precision.

Other experience on query expansion for question answering seems to confirm the positive results that can be obtained by extending the queries with semantically related terms. (Monz, 2003b) reports a certain number of approaches and we refer to his work for a more detailed discussion on query expansion in general. (Monz, 2003b) classified the approaches to query expansion in QA in two major groups:

- the *global expansion*, where knowledge resources, such as WordNet, are used to identify terms that can be added to the query,
- *local expansion*, where additional terms are taken from documents that were retrieved by an initial query that is built from the original questions terms.

Obviously, the type of approach that is most suitable to show the actual utility of language resources is the global expansion method. In particular, the experience that is most important for its inherence with this dissertation is the one reported in (Magnini and Prevete, 2000). Magnini and Prevete describe an experiment carried out in Italian and using ItalWordNet, one of the two Italian lexicons that are the focus of this research. They add to the original query terms their synonyms and morphological variants. Magnini and Prevete (2000) report substantial improvements when using query expansion but also their experience seems not to handle the crucial problem of word sense disambiguation since it is resolved manually by choosing the most appropriate synset before expanding the query.

(Monz, 2003) on the other hand provides an approach which lies between global expansion and predictive annotation[22] and that consists of an expansion performed by exploiting additional terms that are thought to be highly relevant for the question type at hand. For example, the questions type "How many people..?" is expanded with the terms *citizen*, *inhabitant*, *population*, *live*,; the questions asking about the size of something are expanded with the terms *square*, *acre*, *size*, *large* etc. This approach is very interesting because it can be considered a machine-learning alternative to the exploitation of general, not-task-oriented knowledge-sources.

In the following chapters we will provide some experiments concerning the issue of query expansion with terms of ItalWordNet in the Italian QA prototype we built.

2.5.2.3 *SEMANTICS* IN ANSWER EXTRACTION MODULE

After the question analysis and information retrieval passages, Open-Domain QA systems have to face the task of answer identification and extraction from the candidate paragraphs they received from the inner IR module. There is a large range of text features that can be defined to estimate the relevance of a candidate answer for the purpose of answer ranking. Examples are statistical features like term frequency, syntactic dependencies derived through full-text parsing, or semantic features like word senses and their relationships in text and in hierarchical databases like WordNet.

In this last phase, the "semantics" derived from previous analysis has to be exploited in order to pinpoint the answer among the many textual fragments candidate to contain an answer. In particular, the systems can take advantage of the semantic category of the expected answer type. Deciding whether a candidate answer found in a paragraph can be an answer or not requires some level of text understanding, since it is rare that in real texts answers have the form of a simple rephrasing of the question. As we read in (Paşca, 2003):

> *The generative power of natural language makes it extremely difficult to identify automatically which candidate answers are the most relevant, from a (possible very) large set of candidates identified in the passages.*

[22] For predictive annotation we intend a technique consisting in the identification of potential answers in texts by accordingly annotating and indexing them (Prager et al., 2000).

As we did for the first module of the QA architecture, the FALCON system is taken here as an example of a knowledge-based approach to answer identification. In FALCON (Paşca, 2003), two techniques are exploited in order to pinpoint the answer among the many candidates. These techniques are based on recognition of Named-Entities and specific textual patterns.

Named-entity-recognition-based answer identification relies on the fact that during the question analysis the expected answer type has been recognized. Then, the corresponding Named Entity category is searched for among the text tokens in the paragraphs. The candidate answer are then checked to ensure that they do not contain any terms of the question, in order to avoid that a paragraph containing a question term of the same type as the expected answer can be falsely recognized as a possible answer. For example, if the question is *What is the city near Vancouver?* both paragraphs

….Seattle football team reaches the Vancouver stadium in less than a hour by bus…..
…..Vancouver is a nice city near the see….

contain named entities of type CITY>LOCATION and can be retrieved by the search engine, but only the first one contain the answer while in the second one the named entity is the same of the question and has to be discarded. The exploitation of correspondences between expected answer types and named entities is very common, even if differently implemented, in current Open-Domain QA systems. For example, in the PIQAsso system (Attardi *et al*., 2001), among the various filters, we find a semantic filter used to discard paragraphs not containing entities of the expected type. The same approach is also adopted in the DIOGENE QA system (Kouylekov *et al*., 2003).

Pattern-based answer identification, on the other hand, makes use of hand-written patterns that can resolve the case of DEFINITION questions. So, for example, in (Paşca, 2003) a series of patterns is presented, such as (consider AP: Detected candidate answer and QP: phrase to define):

<AP> such as <QP> (*What is autism?* "*developmental disorders such as autism*")
<AP> (also called <QP>) (*What is a bipolar disorder?* "*manic-depressive illness (also called bipolar disorder)*")
<QP> is an <AP> (*What is caffeine?* "*caffeine is an aljkaloid that stimulates*..")
<QP>, a <AP> (*What is a caldera?* "*the Long Valley caldera, a volcanic crater*..")

In FALCON, however, this *knowledge-poor* approach is strengthened with information of a semantic nature: in fact, when possible, the query is expanded with the immediate hyperonym (found in WordNet) of the thing to be defined: if the hyperonym is found in the retrieved passages, it becomes a potential answer. In this way the definition is found in the language resource but has to be supported in the collection too to be considered valid.

2.5.2.4 ENHANCING PERFORMANCE WITH INFERENTIAL CHAINS

The research around the FALCON system has also brought along interesting solutions for WordNet improvement and extension (Harabagiu and Moldovan, 1998, Harabagiu and Moldovan, 2000). The idea was enriching WordNet taxonomy with information on the use of each concept in linguistic context (where context are the glosses of the synsets but also texts derived by journalistic corpora). The extended WordNet (KnowledgeBase WordNet Extended) can be formalized as KBWNextended= (N, R, NG, RG) where N is the current set of nodes representing words or concepts, R represents the existing lexico-semantic relations while NG and RG are the gloss nodes and relations. Moreover, the strategy presented in (Harabagiu and Moldovan, 1998) further enriches WordNet with other information coming from external textual material, giving rise to KBWNcontextualized, that can be defined as a set of (N, R, NG, RG, NC, RC), where NC are context nodes and RC are context relations. The context nodes and relations are organized as a frame where the context nodes are place holders of regular WordNet synsets and the context relations are produced from semantic paths. Next figure (extacted from Harabagiu and Moldovan, 1998), gives an idea of the final configuration of this context-aware WordNet.

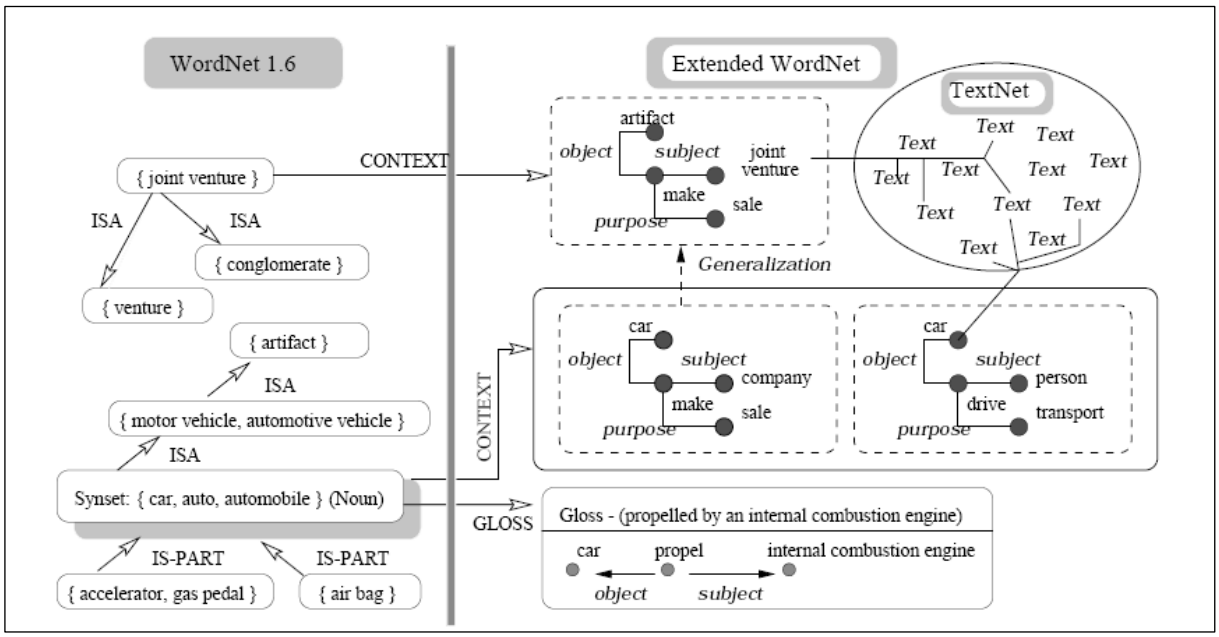


Fig. 12: modules and information types in the Extended WordNet (from Harabagiu and Moldovan, 1998)

In this way WordNet is enhanced in two ways: a) the contextual structures provide with an alternate definition of a concept, targeting automatic processing instead of human understanding, b) the network of words is attached with a web of contextually related texts that is called TextNet.

The first step in Harabagiu and Moldovan's approach consists in extracting information from WordNet glosses. This is due to the exigency of increasing the number of links between WordNet concepts and retrieving the important cross-part of speech connections missing in the American semantic net. Each concept's gloss is transformed into a graph, with concepts as nodes and lexical relations as links. In case of

the synset {interaction}, for example, the gloss "*a mutual or reciprocal action*" is parsed and the following relations are identified:

interaction -- GLOSS→ action -- ATTRIBUTE → mutual
action -- ATTRIBUTE → reciprocal

Obviously, an important factor is represented by semantic disambiguation of the nodes of the graph and (Harabagiu and Moldovan, 1998) presents some of the heuristics that are used to contribute to this task. For example, the fact that the genus term in the definition is already defined as the hyperonym of the concept is sufficient to select the right sense (this is true, for example, for the noun action in the gloss of the concept *interaction*). Other relations among WordNet concepts are exploited to disambiguate the new nodes.

The result is a much richer connectivity between concepts, expressed by means of 13 new relations such as AGENT, OBJECT, PURPOSE, ATTRIBUTE etc. The following picture is presented in (Harabagiu and Moldovan, 1998) to show the new relations acquired by the analysis of the glosses.

Relation	Connects	Concept	Gloss	Example
GLOSS	n_synset–n_synset	{doctor, physician}	(a licensed medical practitioner)	{physician}– –GLOSS→{practitioner}
	v_synset–v_synset	{tease, harass}	(annoy persistently)	{tease}–GLOSS→{annoy}
	adj_synset–adj_synset	{alert}	(vigilantly attentive)	{alert}–GLOSS→{attentive}
	adv_synset–adv_synset	{fully, well}	(completely)	{fully}–GLOSS→ →{completely}
AGENT	v_synset–n_synset	{culture}	(all the knowledge and values shared by society)	{share}–AGENT→{society}
OBJECT	v_synset–n_synset	{glass}	(container for holding liquids)	{hold}–OBJECT→{liquid}
INSTRUMENT	v_synset–n_synset	{chop, hack}	(cut with tool)	{cut}–INSTRUMENT→{tool}
BENEFICIARY	v_synset–n_synset	{ratables}	(property that provides tax income for local government)	{provide}–BENEFICIARY→ →{government}
PURPOSE	v_synset–v_synset	{fork}	(something used to serve or eat)	{use}–PURPOSE→{serve} {use}–PURPOSE→{eat}
ATTRIBUTE	n_synset–adj_synset	{doctor, physician}	(licensed medical practitioner)	{medical practitioner}– –ATTRIBUTE→{licensed}
	v_synset–adv_synset	{browse}	(shop around)	{shop}–ATTRIBUTE→ →{around}
	n_synset–n_synset	{apharesis, aphaeresis}	(omission at the beginning of a word)	{word}–ATTRIBUTE→ →{beginning}
	v_synset–v_synset	{carry}	(move while transporting)	{move}–ATTRIBUTE→ →{transport}

Fig. 13: new relations derived by WordNet glosses (from Harabagiu and Moldovan, 1998)

Thanks to this new set of relations, the glosses are transformed in a network representation like the one presented in (Harabagiu and Moldovan, 1998) as an example (see Fig. 3):

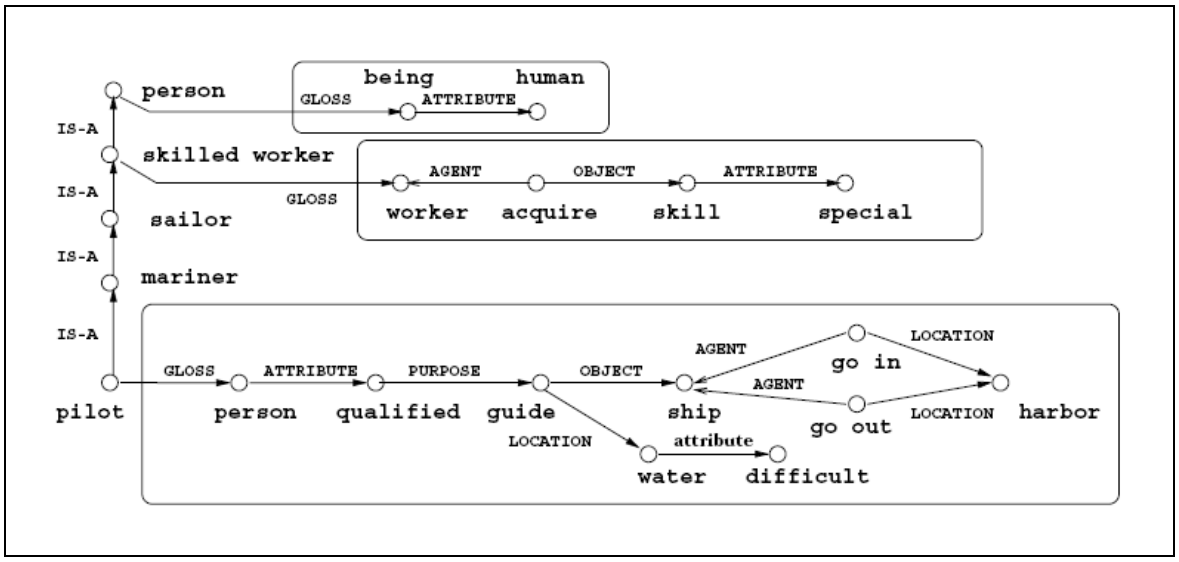


Fig. 14: the graphs resulting from the analysis of the gloss of *pilot* (from Harabagiu and Moldovan, 1998)

The "classic" WordNet relations plus the ones acquired from glosses constitute the units on which another interesting strategy formulated by Harabagiu's group is based on (Moldovan *et al*. 2002 and Harabagiu and Moldovan, 1998): primitive inference rules are implemented as pairs of WordNet semantic relations and from the further combination of primitive rules more complex rules are generated.

In (Harabagiu and Moldovan, 1998) are presented all possible pairs of semantic relations from WordNet1.5 that link three concepts (see Fig. 15).

CONCEPTS : VERB->VERB->VERB

Relations	Number of combinations
IS-A + ENTAIL	8
IS-A + CAUSE-TO	8
ENTAIL + CAUSE-TO	8
IS-A + ANTONYM	4
ENTAIL + ANTONYM	4
CAUSE-TO + ANTONYM	4
SEE-ALSO + IS-A	8
SEE-ALSO + ENTAIL	8
SEE-ALSO + CAUSE-TO	8
SEE-ALSO + ANTONYM	4

CONCEPTS : NOUN->NOUN->NOUN

Relations	Number of combinations
IS-A + IS-PART	8
IS-A + IS-MEMBER	8
IS-A + IS-STUFF	8
IS-PART + IS-MEMBER	8
IS-PART + IS-STUFF	8
IS-STUFF + IS-MEMBER	8
IS-A + ANTONYM	4
IS-PART + ANTONYM	4
IS-MEMBER + ANTONYM	4
IS-PART + ANTONYM	4

CONCEPTS : ADJ->NOUN->NOUN

Relations	Number of combinations
ATTRIBUTE + IS-A	4
ATTRIBUTE + IS-PART	4
ATTRIBUTE + IS-MEMBER	4
ATTRIBUTE + IS-STUFF	4
ATTRIBUTE + ANTONYM	2
PERTAINYM + IS-A	8
PERTAINYM + IS-PART	8
PERTAINYM + IS-STUFF	8
PERTAINYM + IS-MEMBER	8
PERTAINYM + ANTONYM	4
SIMILAR + IS-A	4
SIMILAR + IS-PART	4
SIMILAR + IS-MEMBER	4
SIMILAR + IS-STUFF	4
SIMILAR + ANTONYM	2

CONCEPTS : ADJ->ADJ->ADJ

Relations	Number of combinations
SIMILAR + ANTONYM	2
SEE-ALSO + ANTONYM	3
SIMILAR + SEE-ALSO	4
SEE-ALSO + PERTAINYM	8
SIMILAR + PERTAINYM	4
PERTAINYM + ANTONYM	4

CONCEPTS : ADJ->VERB->VERB

Relations	Number of combinations
PAST-PARTICIPLE + IS-A	8
PAST-PARTICIPLE + ENTAIL	8
PAST-PARTICIPLE + CAUSE-TO	8
PAST-PARTICIPLE + ANTONYM	4

CONCEPTS : ADJ->ADV->ADJ

Relations	Number of combinations
PERTAINYM + SIMILAR	2
PERTAINYM + SEE-ALSO	2

CONCEPTS : ADJ->ADJ->VERB

Relations	Number of combinations
SIMILAR + PAST-PARTICIPLE	

CONCEPTS : ADV->ADJ->VERB

Relations	Number of combinations
PERTAINYM + PAST-PARTICIPLE	4

CONCEPTS : ADV->ADJ->NOUN

Relations	Number of combinations
PERTAINYM + ATTRIBUTE	4

Fig. 15: pairs of relations use for inference rules (from Harabagiu and Moldovan, 1998)

For example, the relation pairs in the first set connect three verb concepts, in the second set three nouns etc. The number to the right of each pair indicates all possible combinations that can be formed with two relations and their inverses. Overall, these pairing result in 314 distinct inference rules.

Other clarifying pictures are the ones we report in Fig. 16 and Fig. 17.

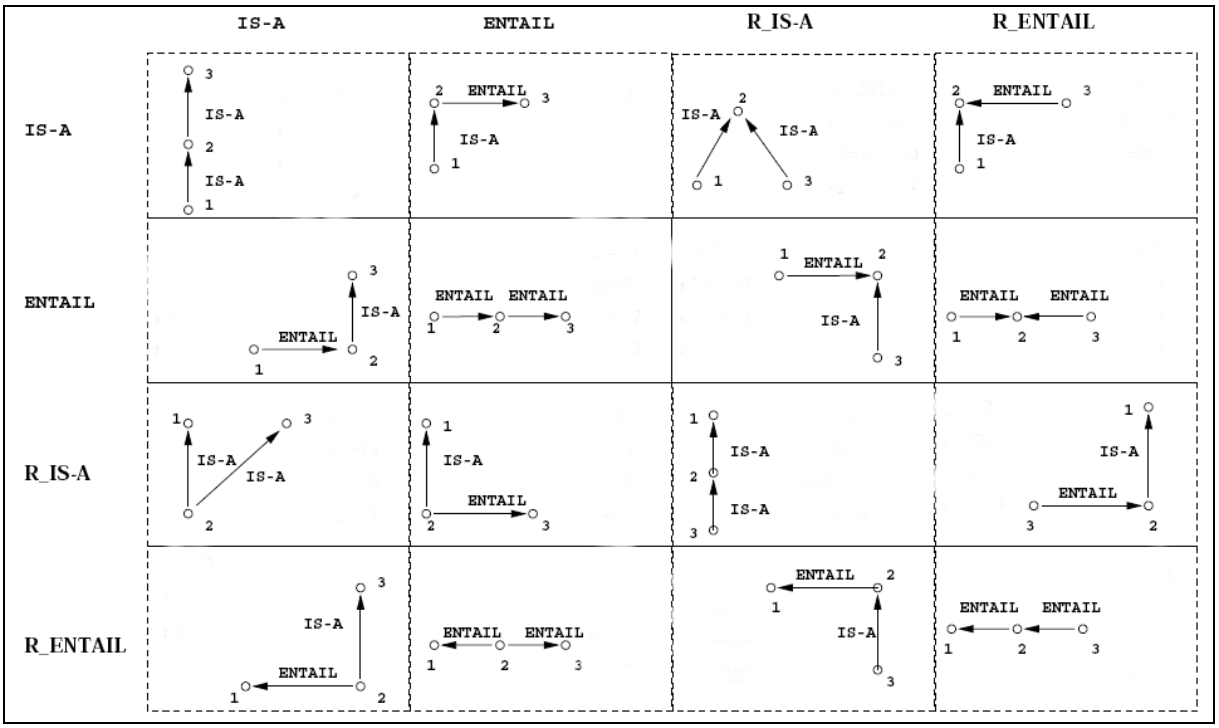


Fig. 16: possible pairs of IS-A and ENTAIL and their reverses (from Harabagiu and Moldovan, 1998)

Rule 1 VC1 IS-A VC2 VC2 IS-A VC3 VC1 IS-A VC3	**Rule 2** VC1 IS-A VC2 VC2 ENTAIL VC3 VC1 ENTAIL VC3	**Rule 3** VC1 IS-A VC2 VC2 R_IS-A VC3 VC1 PLAUSIBLE (not VC3)	**Rule 4** VC1 IS-A VC2 VC2 R_ENTAIL VC3 VC1 EXPLAINS VC3
Rule 5 VC1 ENTAIL VC2 VC2 IS-A VC3 VC1 ENTAIL VC3	**Rule 6** VC1 ENTAIL VC2 VC2 ENTAIL VC3 VC1 ENTAIL VC3	**Rule 7** VC1 ENTAIL VC2 VC2 R_IS-A VC3 VC1 BACKGROUND-OF VC3	**Rule 8** VC1 ENTAIL VC2 VC2 R_ENTAIL VC3 VC1 PLAUSIBLE VC3
Rule 9 VC1 R_IS-A VC2 VC2 IS-A VC3 VC1 PLAUSIBLE (not VC3)	**Rule 10** VC1 R_IS-A VC2 VC2 ENTAIL VC3 VC1 PLAUSIBLE VC3	**Rule 11** VC1 R_IS-A VC2 VC2 R_IS-A VC3 VC1 PLAUSIBLE VC3	**Rule 12** VC1 R_IS-A VC2 VC2 R_ENTAIL VC3 VC1 PLAUSIBLE VC3
Rule 13 VC1 R_ENTAIL VC2 VC2 IS-A VC3 VC1 PLAUSIBLE VC3	**Rule 14** VC1 R_ENTAIL VC2 VC2 ENTAIL VC3 VC1 PLAUSIBLE (not VC3)	**Rule 15** VC1 R_ENTAIL VC2 VC2 R_IS-A VC3 VC1 PLAUSIBLE (not VC3)	**Rule 16** VC1 R_ENTAIL VC2 VC2 R_ENTAIL VC3 VC1 PLAUSIBLE (not VC3)

Fig. 17: inferential rules based on WN relations

The first figure is a graphical representation of all the possible ways the relations IS-A and ENTAIL and can be paired to create inference rules (together with their reverses, REVERSE IS-A and REVERSE ENTAIL).

The second figure presents the resulting inference rules: some of them, thanks to the transitivity of the involved relations, are deductions (for example Rule 1), while other are less certain and provide explanations and background knowledge. (Harabagiu and Moldovan, 1998) presents, as an example of these rules, Rule 4, where VC1 is indicated as a possible explanation for VC3. The two sentences "The criminal apologized" "The criminal confessed his crime" are presented to clarify the explanation relation (cf. Fig. 18) the relations between *apologize* and *confess* in WordNet).

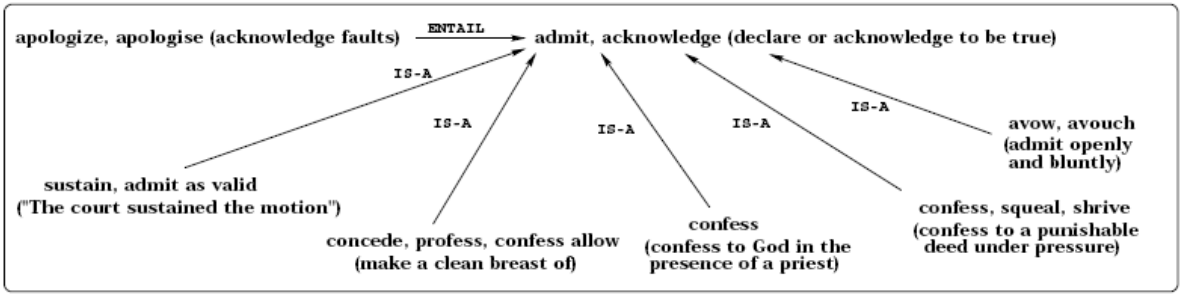


Fig. 18: example of application of Rule 4 (from Harabagiu and Moldovan, 1998)

These primitive rules can be chained by letting the conclusion of one be the premise of the other (for example rules 1 and 2 can be chained without difficulty). An important achievement of Harabagiu and Moldovan's research group is the definition of an algorithm to individuate semantic paths through WordNet concepts. The algorithm is based on a marker propagation method, according to which a marker is placed on a node and it is programmed to propagate from that node only along some selected relations. The input to the algorithm is the semantic knowledge base while the output consists of semantic paths that link pairs of input concepts. (Harabagiu and Moldovan, 1998) also presents an example of the application of their methodology: in case of the two texts *Jim was hungry* and *He opened the refrigerator*, the connection between them can be established by placing markers on the pairs of concepts *hungry - refrigerator* and *hungry – open*. The markers will follow the path traced by their propagation rules.
The resulting paths can be observed in Fig. 19 and Fig. 21 while the inferences generated are monitored respectively in Fig. 22 and Fig. 20.

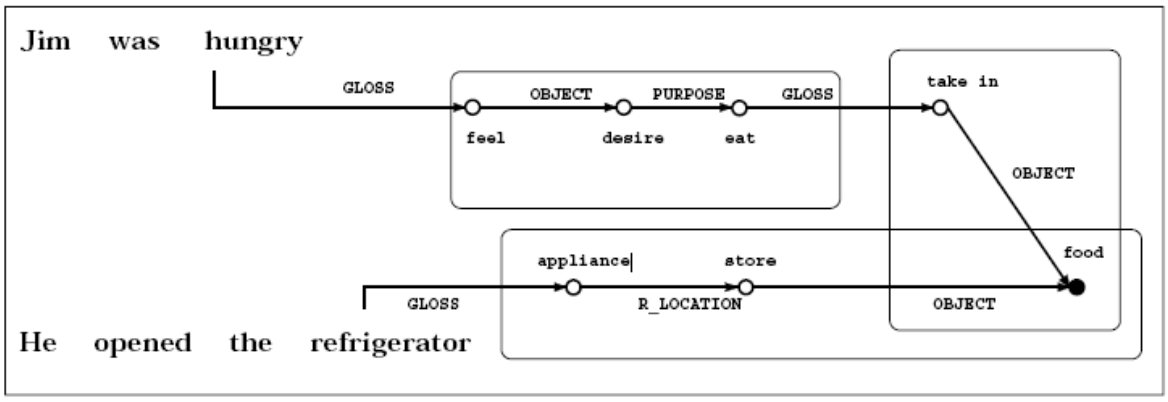


Fig. 19: a valid semantic path from "hungry" to "refrigerator" (from Harabagiu and Moldovan, 1998)

Inference sequence
Jim was hungry.
Jim felt a desire to eat.
Jim felt a desire to take in food.
COLLISION : Jim=he felt a desire to take in food, stored in an appliance, which he opened.
He opened an appliance where food is stored.
He opened the refrigerator.

Fig. 20: inference sequence corresponding to the path in Fig. 19

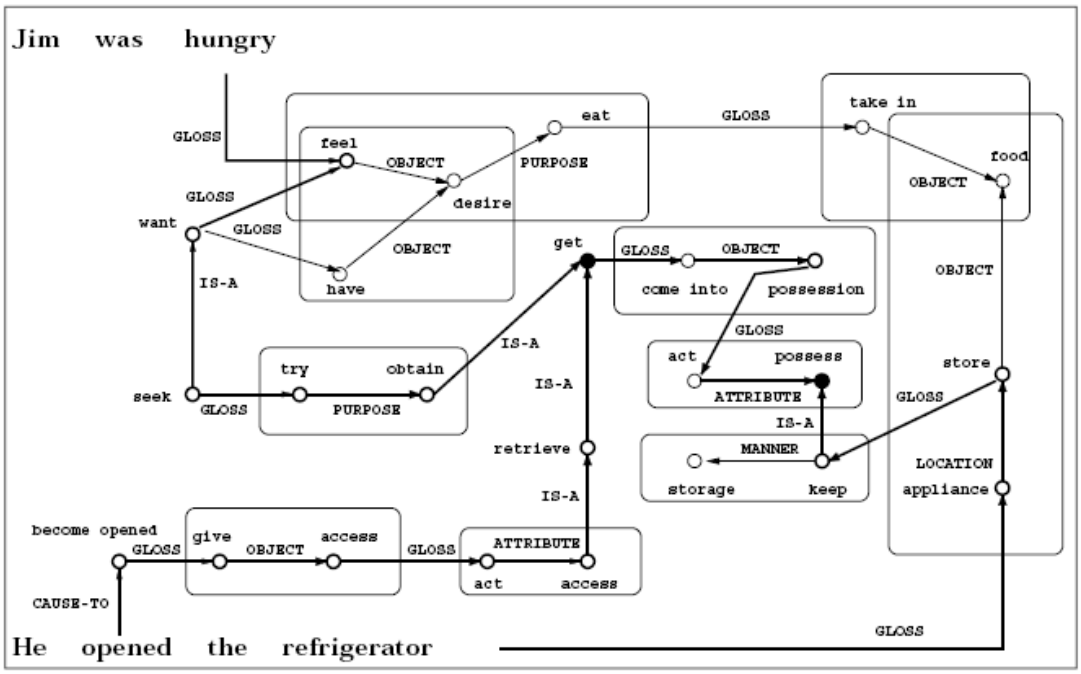


Fig. 21: a semantic path from "hungry" to "to open" (from Harabagiu and Moldovan, 1998)

Inference sequence for path 2	Inference sequence for path 3
Jim was hungry.	**Jim was hungry.**
Jim felt a desire to eat (food).	Jim felt a desire to eat (food).
Jim wanted to eat (food).	Jim wanted to eat (food).
Jim sought to eat (food).	Jim sought to eat (food).
Jim tried to obtain food.	Jim tried to obtain food.
Jim tried to get food.	Jim tried to get food.
COLLISION : Jim=he tried and got food.	Jim came into possession of food.
He retrieved food.	COLLISION : Jim=he came into possession of food.
He accessed food.	He opened the appliance to possess food.
The refrigerator gave him access (to food)	He opened the appliance where food is kept in storage.
The refrigerator became open to him.	He opened the appliance where food is stored.
He opened the refrigerator.	**He opened the refrigerator.**

Fig. 22: inference sequence corresponding to the path in Fig. 21

This strategy open the way to important applicative developments. For Question Answering, the semantic paths can become the ways along which inference move. For example, the paths showed above may become the answers to a question like *Why would someone open a refrigerator?*.

In the light of the experiment carried out on the QA pairs of the questionnaire, is quite obvious how such a strategy would be of great help in the possibility of actually exploiting the semantic relations of our computational lexicons in supporting textual inference. An experience carried out according to the methodology described in (Harabagiu and Moldovan, 1998) is presented in 4.3.7.

3

This second chapter allows us to intuitively enter in issues and problems connected to a question answering task but also to the representation of semantic information. We will analyze the results of a questionnaire to try to understand what type of information makes a text meaningful to people in the specific task of answering a question; then we will verify whether/to what extent the information already available in language resources can be used to support the answer identification process.

3.1 Psycholinguistic Approaches to Question Answering

> *Question answering is a process. If we wish to program a computer to answer questions, we need some sense of what that process looks like. Human question answering is not merely lexical manipulation; the cognitive mechanism used in question answering operate on concept underlying language. The process of question answering must therefore be characterized as manipulation of conceptual information. This thesis presents a process model of question answering as a theory of conceptual information processing.*

The above paragraph is the beginning of the fundamental work of Wendy Lehnert, *The Process of Question Answering* (Lehnert, 1977). After almost thirty years and considering all the changes and innovations, Lehnert's work still represents a point of reference for everyone working in the Question Answering field. What is interesting in Lehnert's approach is that it represents not only an attempt to enable a concrete technology capable of tackling the many challenges of symbolic question answering systems, but also a general psychological model of a cognitive mechanism.

One of the aspects that make QA a difficult subject of study is that, differently from the tasks that received most attention in early stages of AI (playing chess, theorem proving and general problem solving), QA is so *fundamental* and *basic* that it is impossible to introspect, perceive and consciously describe the cognition involved when we answer a question.

Lehnert says that a way to study a cognitive process is to try and describe not only the electrical and chemical activity of neurons in the brain but also the symbolic manipulation underlying that process. The brain itself can be seen as an encoding device that preserves information while a cognitive process is a manipulation that acts on this information. Lehnert's awareness that cognitive processes operate on the meaning of sentences, not on the lexical expression of that meaning, forces us to pursue simulation of human cognition which relies on conceptual representation of information.

As a matter of fact, part of the literature on the psychology of human language understanding is dedicated to the study of mental procedures executed when humans answer questions.
Two different targets can be recognized in this line of research:

i) on one hand, particular effort is dedicated to the definition of a cognitive architecture dedicated to the Question Answering task, (i.e. what cognitive sub-modules are involved when we answer questions? How do they interact with each other? In what order are the operations executed in the brain?). The work presented in Lehnert (1977), Graesser and Murachver (1985), Graesser *et al*. (2002) is dedicated to these "architectural" aspects.

ii) on the other hand, QA constitutes one of the protocols used to inquire about the cognitive processes connected to text comprehension, thus opening the way to some important and more general issues concerning inference generation. The reference point for this work was Graesser *et al*. (1994) and Graesser *et al*. (2001).

Both aspects are useful for our line of reasoning, even if under different perspectives and at different moments: the achievements connected to the first point constituted the psychological basis for the definition of a well-established computational architecture for Question Answering systems, while the study on inferences are more oriented towards the definition of strategies for answer identification.

A surely interesting issue is represented by the recognition of the type of inferences that are generated when a human recognizes a plausible answer to a given question in a text. We thus consider the conclusions of a study by (Graesser *et al*., 2001) dedicated to inference in text comprehension. The paper, instead of starting with text and language and asking what text connections are explicitly articulated, starts with world knowledge and asks what relations are prevalent when we make sense of the world. The aim of Graesser *et al*.'s work is a theory of comprehension that specifies how the meaning representations are constructed on the basis of both world knowledge and the surface linguistic clues.

We will try to go along the same path traced by Graesser and colleagues, taking as input the catalogue of relations they indicated as supporting coherence-based inferences. We will present the results of a questionnaire where parts of those relations are instantiated in text and proposed to human beings in form of question-answer pairs. The aim of the questionnaire is to verify how well human beings are capable of handling complex inferences when they are asked to extract an answer to a given question from a text. The ultimate goal is to verify the possibility of supporting such inferences using information already available in language resources. The results will show that, however, is not always straightforward for humans to match even supposedly banal QA pairs and that lexical mismatches between the texts of question and answer are sometimes a problem not only for machines.

3.2 Empirical approach to QA: the questionnaire

Graesser *et al*. (2002) incorporates a reference literature[23] where empirical data derived from question answering protocols are presented.

[23] We mention here the work presented in (Bransford et al.,1996) and Goldman (1985).

These protocols consist in asking human beings to answer questions about brief stories they read. Questions are factual questions or more complex questions concerning motivations, causal aspects etc.; usually people are asked to speak aloud and motivate their answers, in the attempt to "track" their thoughts.
Moreover, discourse psychologists have explored a number of measures for on-line comprehension processes and inference generation, such as evaluation of the answer time (see for example Robertson *et al*., 1993) and segment fixation times on words during eye-tracking.

3.2.1 Aim, Method and Design of the Experiment

The design of the experiment we propose here is slightly different from previous experiences. We want human beings to test themselves against a task that is very similar to the one that has to be carried out by an automatic QA system. This task is proposed to people in the form of a questionnaire where questions are followed by text paragraphs that can (or cannot) contains an answer[24]. The aim is twofold: on the one hand, the questionnaire is exploited to provide evidence of human capability to handle complex inferences when humans are asked to recognize and to extract an answer to a given question from a text. At the same time, however, we want to analyse the obstacles in recognizing specific question and answer pairs, in order to make emerge the difficulties arising when it is the machine that has to realize the same match.

3.2.1.1 SUBJECTS

Fifty-one people, 28 females and 23 males with an age varying from 24 to 70, participated in the experiment. All the participants were instructed in compiling the questionnaire. The vast majority of the subjects has at least a high school diploma but we also included in the sample a certain number of people less qualified in order to evaluate the role of previous scientific-technical knowledge in answer identification.

3.2.1.2 THE MATERIAL: CREATION OF THE QUESTIONNAIRE

We started from the list of questions used as test set in the CLEF-2004 campaign. The overall list consists of 200 factual questions of the type Chi (Who), Cosa (What), Come (How), Quando (When) and Dove (Where). The answers to the questions were manually searched in the wide newspaper articles collection that constitutes the CLEF-2004 corpus. We then picked the text paragraphs containing the answer to each question, creating in this way a corpus of question-answer pairs.
This corpus was studied and a subset of 21 pairs was selected on the basis of their relevance with respect to initial hypotheses. The idea was to create a sample of pairs where a certain surface distance could be detected between the form of the question and of the paragraph containing the answer. Attention was paid in order to

24 The aim and the methodology of this experiment are very different form the one presented in (Erbach, 2004), where the final aim was to compare the performance of automatic question answering (QA) systems against human QA performance under time constraints.

create question-answer pairs kept together by the relations, surveyed in the study of (Graesser *et al*., 2001), that are supposed to drive coherence-based inferences. So we selected questions and answers that, in our opinion, were linked by means of relations such as CAUSE, HAS-A-PART, IMPLIES, IS-A etc. Moreover, the selection was also based on the presence of phenomena that we consider important under the lexical point of view, such as phenomena that imply the handling of effects of the prototypical nature of meaning, polysemy and figurative shifts of meaning.

The text of question and answer was modified in this way:

1. first of all, we did not want people to already know the answer to the question. As a matter of fact, knowing the answer could have had distorted the results of the test, making it too simple for people to individuate the answer in the depths of the text. For this reason, we changed the name of existing and well-known persons, places, etc. in such a way they could not be recognized. So Arafat become Gifrat, James Bond become Tom Hill and Nelson Mandela become John Mendel. Also the name of the most important Italian newspaper Corriere della Sera was changed to the obviously non-existent Corriere del Nepal, together with changes in the question, that passed from asking the name of the most read newspaper in Italy to asking the most important newspaper in Nepal.
2. The paragraphs constituting candidate answers were changed in order to make them homogeneous in term of length. The idea was that, question complexity being equal, the longer the paragraph, the more difficult the answer.
3. Some changes were introduced in order to test particular hypotheses: for example, the question number 64 of the CLEF 2004 test-set asking "Cosa può causare il tumore ai polmoni?" (What may cause lung cancer?) was changed to "Cosa può causare il tumore all'intestino?" (What may cause intestinal cancer?) and in the paragraph we put colon instead of *polmoni* in order to verify whether the link HAS-A-PART between *intestino* and *colon* can be easily caught by people. In the same way, we invented the question "Chi si è addormentato durante il discorso di inaugurazione dell'anno giudiziario?" (Who did fall asleep during the open address of the "Judical Year"?) because we want to test the capability to catch the implication link between to fall asleep and the verb to snore present in the text paragraph..

We tried to avoid having more than one phenomena of lexical mismatch in the same question-answer pair, this because we would like to keep under control the typology of inference that people make in each sentence.

At the same time, we tried to distribute the various types of relation sought in the question-answer pair in different points of the questionnaire, with the idea that concentrating the same type of connection in the same part of the test could facilitate their recognition and handling.

More than one paragraph can be proposed for the same question: this means that, for example, a question can be followed by three paragraphs, the first containing an answer and the others having nothing to

do with the question. This strategy was adopted to avoid people thinking that an answer can always be found for each question.

For each candidate answer, people were asked to express a number, going for "1" (extremely easy) to "7" (extremely difficult), indicating the difficulty in extracting the answer. A blank space was also available to express doubts or to integrate with comments the answer provided by the subject.

A possible further development of this experiment could be the use of an eye-tracking device to individuate the most meaningful portions of text and of a think aloud protocol in the effort to make inner thoughts emerge.

3.2.2 The questionnaire

In below, we present the QA pairs used in the questionnaire. An overview of the obtained results will be presented in a later section. In this paragraph we provide only the QA pairs where an answer to the question is present while a complete version of the questionnaire will be provided in Appendix A. Providing only the correct QA pairs allows us to more easily focus on the evaluation of the semantic paths connecting the text of the question and of the answer. Moreover, in no case the compreherders selected a completely non pertinent paragraph as candidate answer so we think incorrect QA pairs could be discarded from our discussion without problems. However, we provide some QA pairs selected by compreherders as correct pairs (this even if, in our opinion, they were incorrect, see for example QA pairs number 8, 10, 17).

When analysing the material of the experiment, we will concentrate on phenomena concerning lexical and semantic issues while, obviously, human QA involved decoding at all levels of linguistic description, ranging from recognition of morphological elements of the sentence to syntactic parsing and anaphora resolution. For example, in the first QA pair,:

Q: Dove risiede Gifrat?
A: Il segretario di stato americano incontrerà anche il presidente dell'OLG Hibraim Gifrat, il quale ha preso da pochi giorni residenza permanente nella striscia di Gaza.

among several other decoding operations, the comprehender has to establish a link between the name *Gifrat* in the question and the pronoun *il quale* in the answer, passing through the anaphoric link between Gifrat and *il quale* in the answer text.

Moreover, we will not discuss the mapping between the interrogative elements at the beginning of the question and the type of the expected answer. The references on psychological issues connected to QA (Lehnert, 1977, Graesser and Murachver, 1985) show that when the comprehender reads *Dove* at the beginning of the question, he/she immediately looks for an answer of the type location, if he/she reads *Quando* the sought answer is an expression of time etc. This mapping between question stem and expected answer type is very important both for psychological and computational issues but in this discussion on the questionnaire we will deal with it only when the question stem is of the type *Quale* and *Che*, since it implies the decoding of the semantics of the lexical unit modified by the stem (*quale città*, *quale persona* etc.).

Thus in general, given the aim of our work, we will concentrate our attention only on phenomena that can mainly be characterized as pertaining to lexico-semantic issues.

1. Dove risiede Gifrat?	*Il segretario di stato americano incontrerà anche il presidente dell'OLG Hibraim Gifrat, il quale ha preso da pochi giorni residenza permanente nella striscia di Gaza.*

2. In che giorno è stato ucciso Aldo Moro?	*Aldo Moro è stato rapito il 2 febbraio del 1978 e la sua morte, il 9 maggio del '78, ha sconvolto l'Italia, gettando nel panico cittadini ed istituzioni.*

3. Qual è la professione di Tom Hill?	*Da allora uscirono altri quindici film, tredici dei quali hanno come protagonista Tom Hill, agente segreto della CIA.*

4. Cosa può causare il tumore all'intestino?	*E' una zona in cui l'aria e' irrespirabile, non dimentichiamo che i genovesi sono ai primi posti per morte di tumore.*

5. Cosa può causare il tumore all'intestino?	*Ricercatori giapponesi sostengono, dopo accurati studi, che gli scarichi diesel causano il tumore al colon.*

6. Quali esseri viventi sono in grado di assorbire l'anidride carbonica?	*Hajime Kayane sostiene che le barriere coralline presenti nel mondo sono oggi in grado di assorbire il 2 per cento delle emissioni di anidride carbonica nel mondo intero..*

7. Qual è la capitale del Bhutan?	*Lo scorso 24 ottobre, durante il quindicesimo round di colloqui a livello ministeriale,le due nazioni avevano sottoscritto nella capitale bhutanese Thimpu un accordo bilaterale.*

8. Quale animale tuba?	*I gincorli, che da qualche anno sono arrivati dai Balcani, hanno incominciato a tubare prima del tempo..*

9. Quale animale tuba?	*Il musicista aveva ritratto, con pari esattezza visiva, il ruggito del leone, il cinguettio dell'usignolo e il tubare dei colombi.*

10. Quale animale tuba?	*I fidanzati tubavano sulle panchine, sussurrando dolci parole d'amore all'ombra degli alberi, giurandosi eterna e reciproca fedeltà.*

11. Quando e' stato stipulato il Trattato di Maastricht?	*I commentatori hanno parlato a lungo della ratifica del Trattato di Maastricht avvenuta nell'autunno del 1992.*

12. Quanti membri della scorta sono	*la strage di Capaci, dove morirono il giudice Giovanni*

morti nell'attentato al giudice Falcone?	*Falcone, la sua compagna Francesca Morvillo e tre degli agenti di scorta..*

13. Quanti anni di prigionia ha subito John Mendel?	*John Mendel ha compiuto oggi una visita carica di dolorosi ricordi nel penitenziario di Robben Island dove egli subì 19 dei 27 anni di carcere.*

14. Di quale nazionalità erano le petroliere che hanno causato la catastrofe ecologica vicino a Trinidad e Tobago nel 1979?	*Al largo di Trinidad e Tobago (Mar dei Caraibi), entrano in collisione le navi "Atlantic Express" e "Aegean Captain", ambedue battenti bandiera liberiana.*

15. Quali esseri viventi sono in grado di assorbire l'anidride carbonica?	*Secondo i meteorologi, i coralli sono in grado di assorbire CO2 e altri gas responsabili dell'incremento della temperatura del pianeta..*

16. Come vengono chiamati i piloti suicidi giapponesi?	*Nella battaglia di Okinawa morirono più di mille piloti kamikaze che si gettarono sulle posizioni nemiche con gli aerei imbottiti di esplosivo e muniti della benzina sufficiente solo per il viaggio di andata..*

17.Come vengono chiamati i piloti suicidi giapponesi?	*Il kamikaze giapponese, tanto bravo e veloce quanto sprovveduto, rompendo dopo pochi minuti il motore Yamaha della sua vettura ha inondato tutta la pista d'olio in maniera tale che si sarebbe potuta preparare un'insalata.*

18. Che scuola frequenterà William, il figlio maggiore del principe Carlo?	*Magliette con l'immagine del principe William vestito con il tradizionale abito a code degli studenti di Eton sono state ritirate dal commercio in seguito ad una protesta di Buckingham Palace.*

19. Quando e' stato stipulato il Trattato di Maastricht?	*La conclusione del Trattato di Maastricht è del 1991, anno ricco di avvenimenti importanti per l'Europa intera.*

20. Quanti membri della scorta sono morti nell'attentato al giudice Falcone?	*Nel secondo anniversario dell'attentato di Capaci, vengono ricordati il giudice Giovanni Falcone e gli agenti di scorta Antonio Montinari, Vito Schifani e Rocco di Cillo.*

21. Qual è il quotidiano nepalese più letto?	*Fatturato complessivo in lieve calo per la stampa nepalese nel 1993 mentre il Corriere del Nepal si conferma al primo posto nella classifica dei quotidiani nazionali.*

22. Chi si è addormentato durante il discorso di inaugurazione dell'anno giudiziario?	*durante il discorso di inaugurazione dell'anno giudiziario il presidente del senato stava russando, con evidente imbarazzo del resto della platea.*

23. In che giorno è stato ucciso Aldo Moro?	*Aldo Moro è morto il 9 maggio 1978, tre mesi dopo il suo sequestro ad opera delle Brigate Rosse.*

24. Dove risiede Gifrat?	*Immagini inconsuete scuotono la coscienza di Israele e pongono domande difficili, domande che, sicuramente, a casa sua, a Gaza, Gifrat si pone per converso.*

25. Cosa può causare il tumore all'intestino?	*Studi recenti dimostrano come gli OGM causino il cancro all'intestino.*

3.2.3 The questionnaire: discussing the results

The following two diagrams allow us to study: i) what percentage of subject correctly answered each question and ii) how difficult it was considered to extract the right answer.

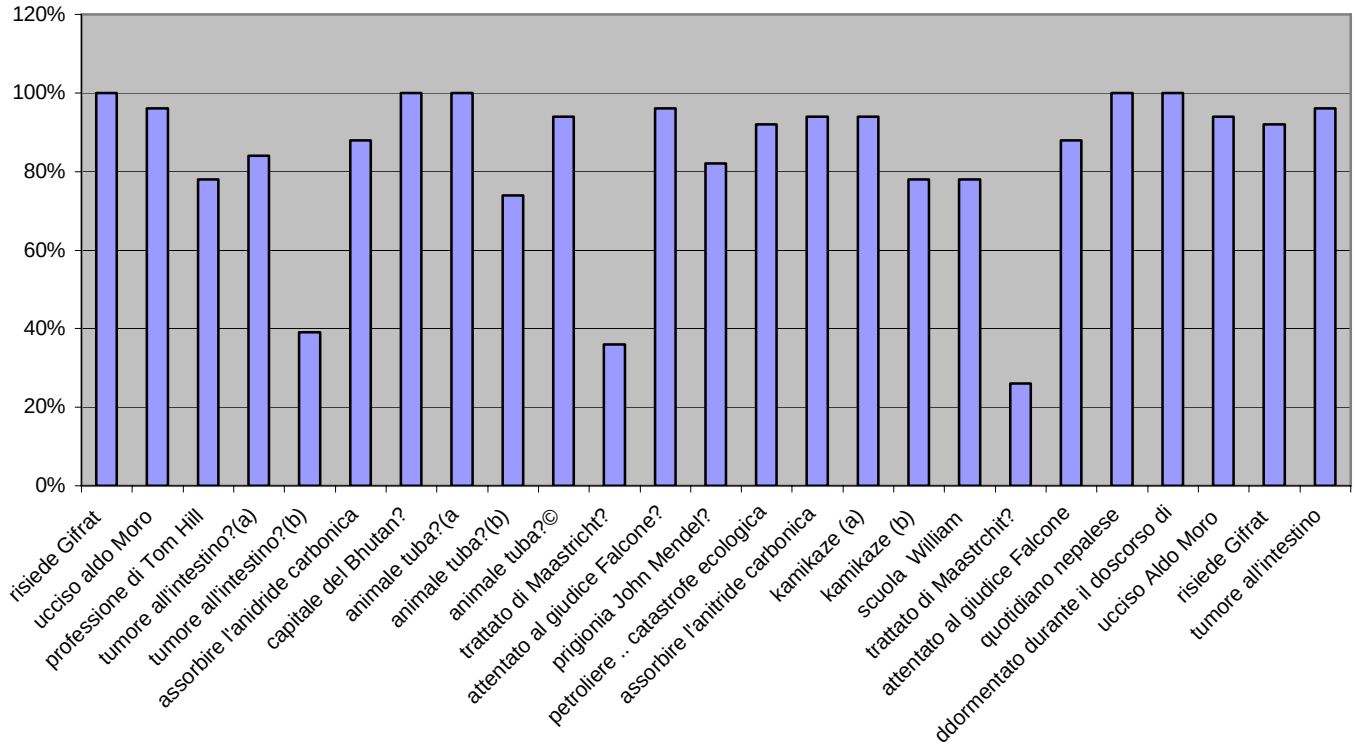


Fig. 23: percentage of right answer for each QA pair

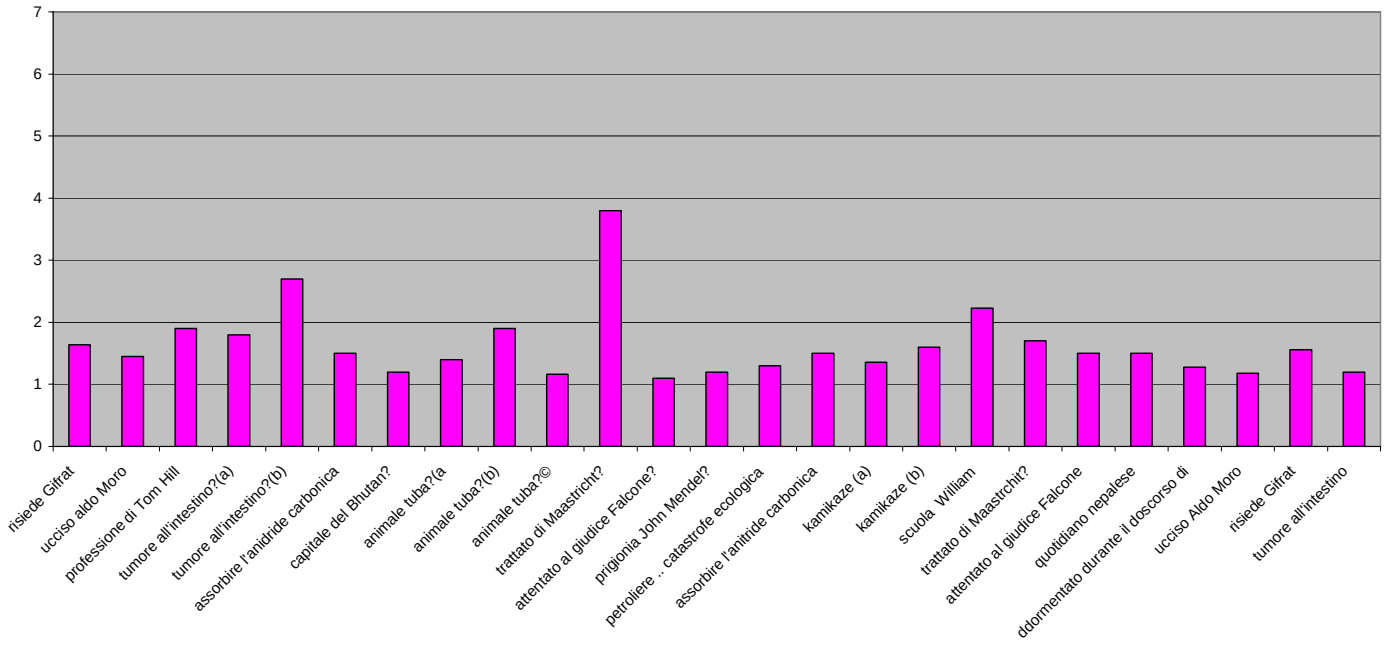


Fig. 24: average of expressed complexity for each QA pair

In general, the questionnaire confirms human capability of performing "information extraction" on textual material, establishing implicit connections, picking suggestions from the context and exploiting previous knowledge they have. We will see, however, that some important exceptions can be identified.

The two diagrams have to be analyzed together since their results are strongly interconnected. As a tendency, "easiest" QA pairs to answer are also the ones with the highest percentage of correct answers. This is true, for example, for the first QA pair, where the lexical mismatch between *risiedere* and *(prendere la) residenza* was easily handled by all the subjects. Furthermore, the last QA pair, still concerning the question "Dove risiede Gifrat?", was effectively handled together with the mismatch between *casa* and *risiedere* (even if with a little more perceived complexity maybe because of the morphological difference). The interpretation of the entailment link between the verbs *russare* (to snore) and *addormentarsi* (to get asleep) of QA pair#22 was evidently effortless. Also some pairs involving decoding at pragmatic-world-knowledge level were very easily handled, such as the QApair#21 (with the easy interpretation of "essere al primo posto nella classifica" as an equivalent of "essere il più letto"). Two of the QA pairs with the highest number of correct answers and the lowest perceived complexity (respectively 96% and 94%, 1,45 and 1,1) are the number 2 and 23 pairs: the date of Aldo Moro's death was correctly detected and extracted, demonstrating that the connections between uccidere and morire on one side and uccidere and morte on the other can be easily established. At the same time, however, it is quite interesting to note that for almost 6% of the subjects, the expected answer for a question asking about a day is just the day of the month, not the complete date (since they answered *9 maggio* and not *9 maggio 1978*). Is has to be said that *giorno* in this sense can be ambiguous and that even the name of the day of the week would not have constituted an actual invalid answer.

In the case of the two QA pairs asking what living entities are able to absorb carbon dioxide, people do not seem to have met any problems handling a non-prototypical sense of living entity: this is what emerges both from the analysis of the percentage of correct answers (88% and 94%) and by the declared perceived complexity (1,5 and 1,3). What is strange is that i) difficulties seem even lower when the carbon dioxide was indicated with its symbol ii) many subject felt the need to ask, in the comment field, whether coral reef is really a living entity.

Notwithstanding the general positive results obtained by the subjects of the experiment, sometimes the results are not as good as we could have expected. For example, in QApair#10 (dedicated to "animals that coo"), we wanted to evaluate the comprehender's capability to correctly handle the various senses of the word *tubare*. The human subject of the verb selects, in this case, the figurative sense of tubare, referring to the soft speaking of lovers. But, quite surprisingly, only 74% of the subjects answered that the sense of *tubare* (to coo) in the candidate answer was not the one present in the question but rather a figurative sense: 25% of the subjects answered "fidanzati" to this question, saying in the comment field that lovers are animals as well!

The interpretation of the noun *protagonista* in QApair#3 led 18% of the subjects to answer "attore" to the question "Qual è la professione di Tom Hill?" even if that word was not present in the text. As a matter of fact, the text of the paragraph is somehow ambiguous: the noun *protagonista* is polysemous and, in order to answer the question, the comprehender has to resolve this ambiguity.

On of the QA pairs with the highest complexity is the number 18, the one asking what school Prince William is going to attend: 78% of the subjects answered Eton but expressed perplexity in the comment field saying that they could not be sure and that the answer is only probable.

In the first of the three QA pairs asking what can cause intestinal cancer, 84% answered "scarichi diesel", demonstrating in this way that they were able to identify the *colon* as a part of the *intestine*. A smaller percentage of subjects did not extract the answer and some of them explicitly asked in the comment field if the colon was a part of the intestine, because they were not sure or did not know. The second QA pair dedicated to causes of intestinal cancer was judged incorrect by more than 60% of the subjects: the text of the candidate answer was too generic (it deals with causes of cancer in general) while the question specifically asks about causes of intestinal cancer. What is interesting is that for the majority of the subjects the "logical" true consequence that states that what is valid for a larger class is also valid for its sub-classes does not hold. The last QA pair dedicated to this question was correctly answered by more than 96% of the subjects. Someone, however, sustained that cancer and tumour are not true synonyms, and did not find correct extract that answer.

The couple of QA pairs concerning the stipulation of the Maastricht Treaty were not correctly handled: for the vast majority of the subjects (respectively 64% and 74%) the meaning of *stipula* coincides with the notion of *ratifica* and *conclusione*. These two QA pairs are among the ones with the lowest percentage of correct answers and, in the case of number 11, the highest perceived complexity. What we wanted to test with these pairs was the fact that probably "common sense" will bring people to associate the

meaning of the two words, which under a technical point of view are very different. This QA pair belongs to the CLEF-2003 campaign and it was not changed in the questionnaire. The international committee that evaluated the results of the systems judged this answer a valid one; we disagree with that opinion but it is not really important, what is really useful is to evaluate the fact that the majority of the subjects involved answering this question in the questionnaire implicitly expressed how close the two meanings (*ratifica* and *stipula*) are in their opinion.

Another case that evidently caused problems to the subjects of the experiment is QApair#17: about 20% of the subjects seemed to think that the bikers which race for Yamaha can be considered suicide pilots too.

In another case, subjects performed "too well", identifying the answer "gincorlo" in the QApair#8. Nothing in the answer explicitly said that the gincorlo was a type of animal and a lot of subjects felt the need to comment their answer saying that they did not know that the gincorlo existed but that all the things in the paragraph suggested that the gincorlo was an animal (in particular the "arrivare dai Balcani" that was interpreted as a migration). What is interesting in this case is that to answer the question people did not have to access a kind of animal taxonomy inside their brain (in that case they would not have found any "entries" for the gincorlo) but rather exploit some hints from the context, and they do not seem to have had any problems with this.

Here what keeps the question answer pair together is the link that could be established between the word professione and the noun agente (or agente segreto della CIA). The text of the paragraph is also interesting because it is somehow ambiguous: the noun *protagonista* is polysemous and, in order to answer the question, the comprehender has to resolve this ambiguity. We expect a certain number of people to think Tom Hill is the name of the actor who plays the secret agent. We think that, even if the ambiguity could represent a difficulty, we think that the right answer for this question should be *agente segreto*.

3.3 From human knowledge to lexical-semantic language resources.

In the previous paragraphs, we showed how humans, in the effort to pinpoint an answer to a given question in a text, seem to be able to effortlessly bridge the gap between distant surface textual units. The inferences built in order to overcome the lack of explicit connections are made on the basis of knowledge (world-knowledge, lexical knowledge, pragmatic knowledge, encyclopaedic knowledge)[25]; now we want to investigate whether the information stored in the lexical databases we are working with, ItalWordNet and SIMPLE-CLIPS, can constitute a source of the same knowledge people use when making inferences.

In this phase our analysis is intended only to verify whether explicit connections potentially able to support inference are available in our lexicons; a more serious and difficult issue consists in establishing whether such connections can be really exploited in a concrete system, i.e. whether our lexicons are really

[25] It is clear that the difference among these classes is not sharp and it is not easy to really understand which is the distinction between, for example, lexical and encyclopaedic knowledge.

computable in many of their parts. For the best of our knowledge, only one work is dedicated to the full exploitation of the range of links and information available in WordNet as support to inference (the work presented in Harabagiu and Moldovan, 1998 and Harabagiu and Moldovan, 2000 and described in 2.5.2.4).

Moreover, the presence of potentially useful connections is not enough and has to be supplied with other methods to individuate potentially answers. This means that, for example, to individuate the answer in QA pair #2:

Q: *In che giorno è stato ucciso Aldo Moro?*
A: *Aldo Moro è stato rapito il 2 febbraio del 1978 e la sua morte, il 9 maggio del '78, ha sconvolto l'Italia, gettando nel panico cittadini ed istituzioni.*

the system will have first to extract the paragraph containing the answer (for example by expanding the query term *uccidere* with the term *morte*) and then rely on a Named Entity Recognizer to extract the date *9 maggio del '78*. Other strategies can be exploited (for example based on syntactic rules) and we will advance on the different methods in the next chapter.

More precise information is needed: the useful connections may not be in the language resources for two reasons: because their linguistic models do not support the representation of the required information or because that specific information has not been covered (yet) by the encoding process. In what follows we will explicitly indicate both the cases.

3.3.1 Bridging the gap between Question and Answer: contribution of LRs

We again present all the QA pairs, providing, this time, the paths that, in the two language resources, support the matching between the forms in the question and in the candidate answer. The exemplification is first provided for the ItalWordNet semantic net and then for SIMPLE-CLIPS.

QA pair #1:
Q: Dove risiede Gifrat?
A: *Il segretario di stato americano incontrerà anche il presidente dell'OLG Hibraim Gifrat, il quale ha preso da pochi giorni residenza permanente nella striscia di Gaza.*

In the first QA pair, we can see that, by exploring the IWN branches, we discover a path which connects the verb *risiedere* and the noun *residenza*. The connection is not direct but exploits an intermediate node, the synset {casa, abitazione, …, dimora}. The relation connecting the verb with the noun concept is of the type INVOLVED/ROLE_LOCATION.

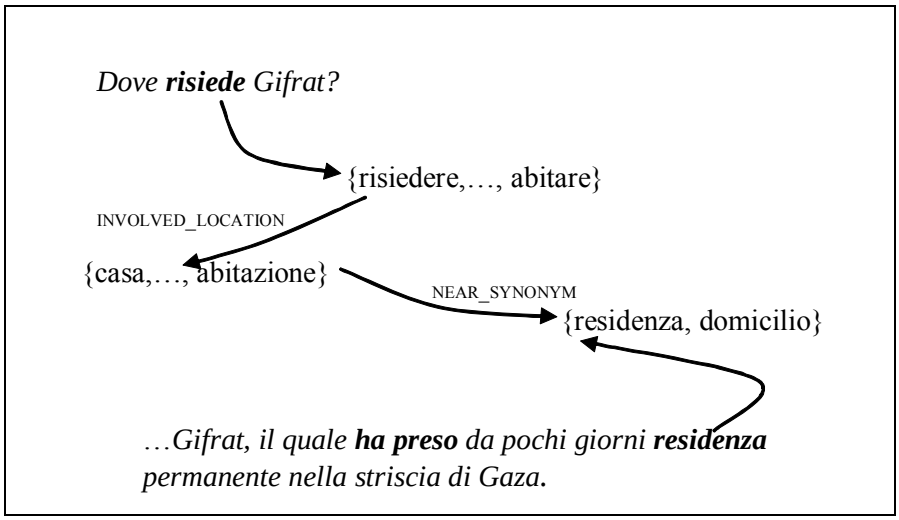


Fig. 25: IWN nodes and links between *risiedere* and *residenza*

This relation is a subtype of the more general INVOLVED/ROLE relation, which can be used to connect different ontological types, more specifically, the different roles and functions that 1st and 3rdOrderEntities may have in events (2ndOrderEntities). From a cognitive point of view, function seems to be one of the major features that organizes human knowledge and functionality is widely reflected in the lexicon and could be useful for language engineering tasks. Functional relations are often related to telicity but, since they also cover other aspects of semantic entailment, in the EuroWordNet project they were referred to as more generic involvement relations. If the relation goes from a concrete or mental entity (only nouns denoting 1st or 3rdOrderEntities) to verbs or event denoting nouns (2ndOrderEntities), it will be called ROLE, the inverse from events (2ndOrderEntities) to concrete or mental entities (nouns) is called INVOLVED. In the EWN documentation (cf. for example Vossen, 1999), we read also that ROLE/INVOLVED relations should not be confused with a way to express true *selectional restrictions*. For example, we should encode a relation of the type INVOLVED_INSTRUMENT between the verb *to hammer* and the instrument *hammer*, since it is conceptually salient and will immediately be triggered regardless of the context. Nevertheless, this information should not be interpreted as expressing a selectional restriction since the instrument of *to hammer* can be any physical objects and not only a *hammer*. The subtype ROLE/INVOLVED_LOCATION is used when the encoder wants to link a place with the noun or verb denoting the action that happens in that place (for example, *school* and the *teaching* activity).

In the case of the first QA pair, the intermediate node (casa,…, abitazione} is also linked, by means of a NEAR_SYNONYM relation, to the target concept *residenza*. The NEAR_SYNONYM relation was exploited in EWN/IWN when a close relation between words could be detected but not sufficient to make them members of the same synset.

In SIMPLE-CLIPS, nothing seems to link the *risiedere* and the *residenza* concepts. *Risiedere* is classified under the Stative_Location Type and linked, by means of an ISA relation, to the concept *abitare* (that in the IWN database is indicated as a synonym of *risiedere*). This lexical entry is also liked to a lexical

predicate consisting of two arguments, respectively the Role ProtoAgent (that can be a Human), and the Role Location (that has to be a Concrete Entity).
The Semantic Unit *residenza* is classified as a Geopolitical Location having as hyperonym the SemU *luogo*.

QA pair #2:
Q: *In che giorno è stato ucciso Aldo Moro?*
A: *Aldo Moro è stato rapito il 2 febbraio del 1978 e la sua morte, il 9 maggio del '78, ha sconvolto l'Italia, gettando nel panico cittadini ed istituzioni.*

What we have to look for in our LRs is a path connecting the verb *uccidere* (to kill) in the question and the noun *morte* (death) in the answer. Moreover, the answer type term *giorno* (day) has to be classified in order to immediately trigger the search of a temporal expression.
In IWN, we see that a certain path can be traced connecting *uccidere* and *morte*, exploiting the intermediate node *morire*: as a matter of fact, in the LR it is stated that that *uccidere* (to kill) results in *morire* (to die) and that *morte* (death) is equivalent to *morire* but belongs to a different part of speech.

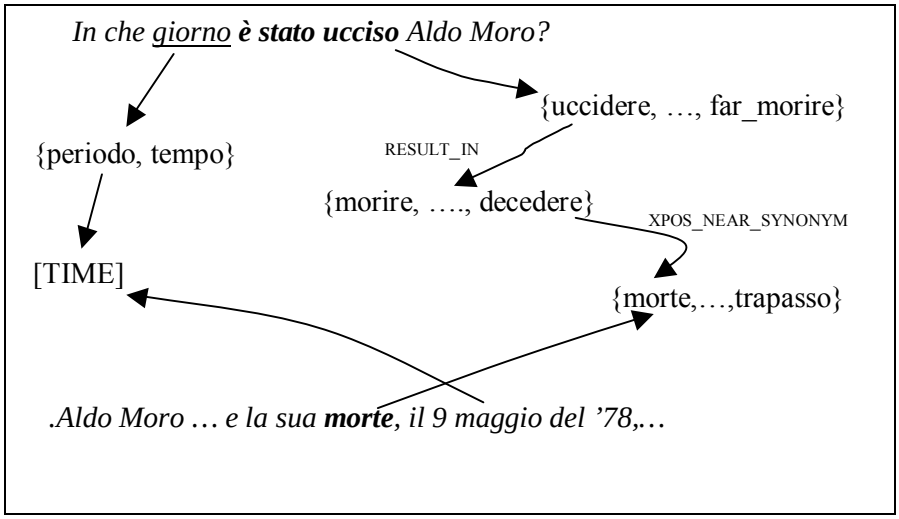


Fig. 26: IWN nodes and links between *uccidere* and *morte* and derivation of the expected answer type

This configuration of concepts is realized by using specific semantic relations available for encoders in ItalWordNet: the RESULT_IN relation, which is a subtype of the more general CAUSE relation, and the XPOS_NEAR_SYNONYM relation.

In EuroWordNet, the CAUSE relation is used to express causativity and to link 2ndOrderEntities. In this sense the relation is thus type-persistent but can apply across POSs.

The other connection, the one between *morte* e *morire*, is expressed using the XPOS_NEAR_SYNONYM relation used to establish (often derivational) links between near synonyms belonging to different parts of speech and referring to situations and events (2ndOrderEntities). Obviously, it would have also been possible to encode a relation of the type RESULT_IN directly targeting the nominal concept *morte*.

The other type of information human beings seem to effortlessly handle is the derivation of the expected answer type from the interpretation of the noun modified by the interrogative adjective *Quale*. In this specific case, we can infer that what we have to look for in the candidate answer is a temporal expression, since the noun *giorno* is correctly classified under the Top Concept TIME.

Exploiting the SIMPLE database we can obtain the same result: the Semantic Unit *uccidere* is indirectly connected with *morte* via the intermediate adjectival Semantic Unit *morto* and the verbal Semantic Unit *morire*. Also *giorno* is correctly interpreted as requiring a temporal expression.

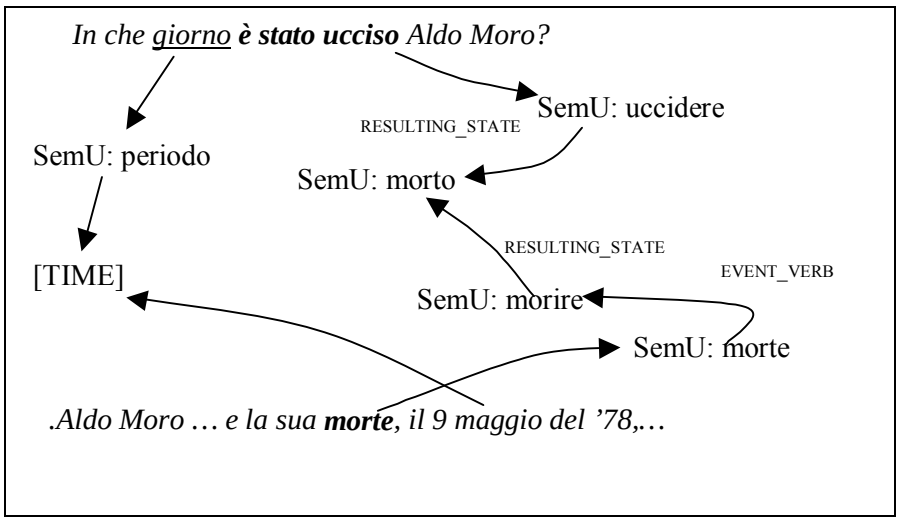


Fig. 27: SIMPLE-CLIPS: arches and nodes connecting uccidere and morte and expected answer type

In SIMPLE-CLIPS the Resulting_state is a relation of the Constitutive that allows the encoder to establish a link between a SemU expressing a transition and a SemU that is supposed to be the resulting state of the transition. The link between *morte* and *morire* is instead of the type EventVerb, used to link an event nominal, equivalent to the event expressed by the verb.

QA pair #3:

Q: *Qual è la professione di Tom Hill?*

A: *Da allora uscirono altri quindici film, tredici dei quali hanno come protagonista Tom Hill, agente segreto della CIA*

Here what keeps together the question answer pair is the link that could be established between the nouns *professione* and *agente* (or *agente segreto della CIA*).

In IWN, the path that connects the two concepts is quite complex and marked by the hyperonymy and role_agent relation:

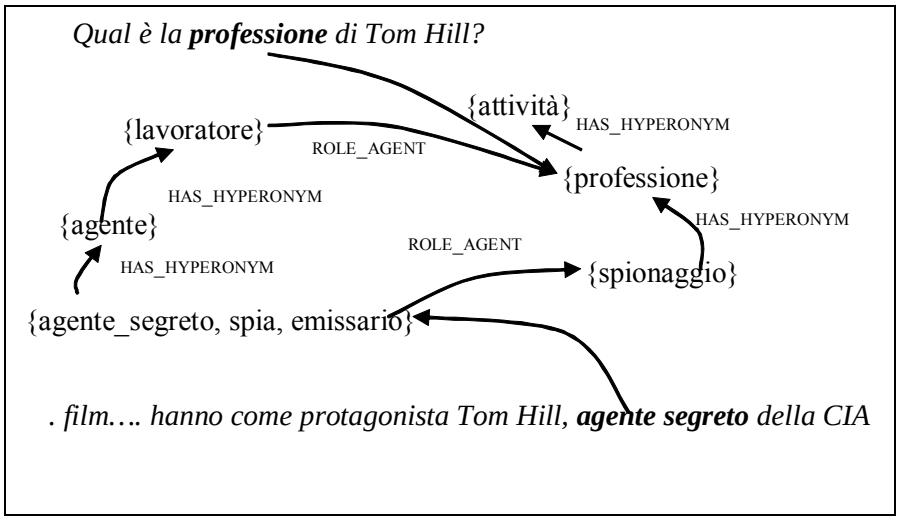


Fig. 28: IWN relations connecting *professione* and *agente segreto*

The two taxonomies of i) humans performing a job and ii) the job and activities, are not completely distinct (as they are in Princeton WordNet): they are connected in many points by the ROLE/INVOLVED_AGENT, one of the many sub-types of the ROLE/INVOLVED relation, specifically used to express the thematic role of agent of activity.

In SIMPLE-CLIPS, we cannot find any common point between the SemU *agente* (*agente segreto*, being a multiword, is not present in the database) and the SemU professione: a*gente* has an ISA relation with the noun *militare* and also two relations of the type is_the_activity_of respectively with *indagare* (to investigate) and *difendere* (to defend). *Professione*, on the contrary, has only an ISA relation with *attività* (activity). The two SemUs do not share the same Semantic Type but what is interesting is that *agente* belongs to the Profession Template: in this case the key information that should be exploited would't be the lexical one interpreted by the SemUs but the Ontological information.

Moreover, the Semantic Type Profession is in SIMPLE-CLIPS one that exploits the Unified Types, i.e. types where the agentive and/or telic multiple coordinates inherently characterize the essence of that type. As a matter of fact, in SIMPLE-CLIPS types of different complexity are envisaged: while some types are simple, i.e. monodimensional, others are inherently defined by the agentive and/or telic dimension they include. While monodimensional relations provide an exhaustive characterization in the case of a type like [Animal], for the [Profession] type it is not enough since it needs extra coordinates, which refer to the functional dimension it incorporates.

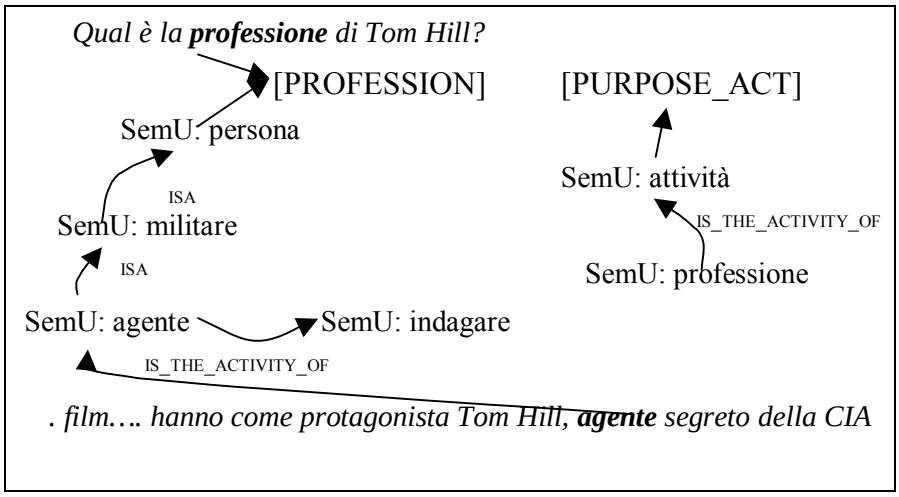


Fig. 29: *agente* and *professione* are not connected but the Semantic Type plays an important role

QA pair #4:
Q: *Cosa può causare il tumore all'intestino?*
A: *E' una zona in cui l'aria e' irrespirabile, non dimentichiamo che i genovesi sono ai primi posti per morte di tumore.*

QA pair #5:
Q: *Cosa può causare il tumore all'intestino?*
A: *Ricercatori giapponesi sostengono, dopo accurati studi, che gli scarichi diesel causano il tumore al colon.*

Both in IWN and in SIMPLE-CLIPS the synecdoche between the *colon* and the *intestino* body parts can be expressed by means of a meronomy relation, instantiated in the two lexicons respectively as the HAS_MERO_PART and has_as_part relations[26]. What is different is that while in SIMPLE-CLIPS *colon* and *intestino* are directly connected, in IWN an additional meronymy link has to be followed:

[26] In both lexicons, many other types of meronymy relations are available.

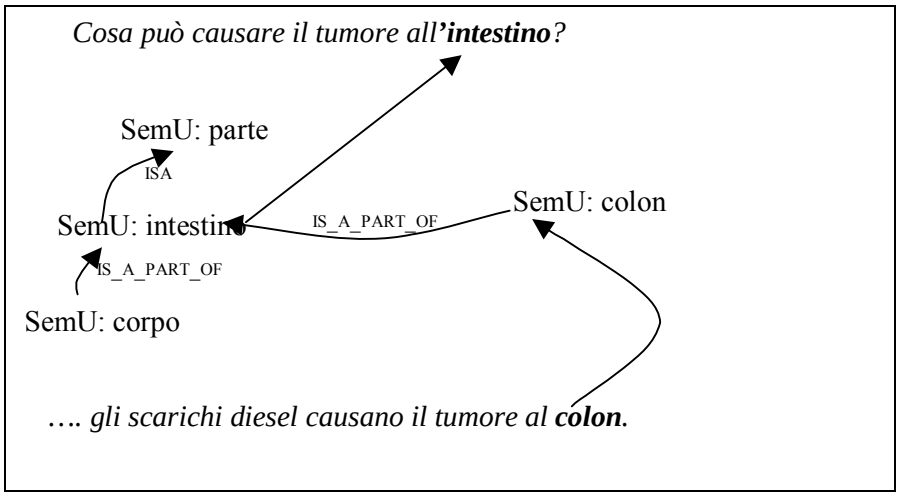


Fig. 30: semantic relations directly linking colon and instestino in SIMPLE-CLIPS

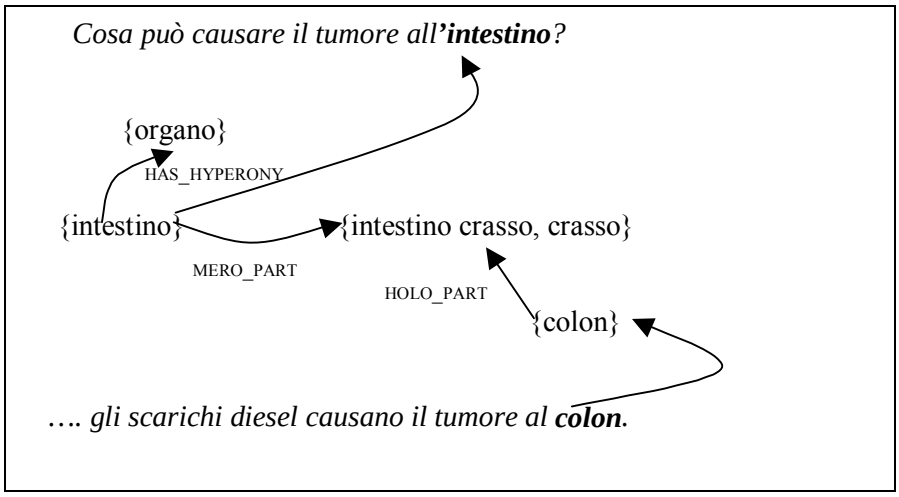


Fig. 31: semantic relations indirectly linking colon and instestino in IWN

QA pair #6:
Q: *Quali esseri viventi sono in grado di assorbire l'anidride carbonica?*
A: *Hajime Kayane sostiene che le barriere coralline presenti nel mondo sono oggi in grado di assorbire il 2 per cento delle emissioni di anidride carbonica nel mondo intero...*

Neither in IWN nor in SIMPLE-CLIPS were we able to find a suitable path that could help to identify the coral reef as a living entity. Neither in IWN nor in SIMPLE-CLIPS coral reef were present as a multiword lexical entry. Moreover, IWN categorizes *barriera* (reef) as a geological formation (under the ontological Top Concepts Place and Substance). C*orallo* (coral) is defined as an animal, thus as a living entity. SIMPLE-CLIPS does not have this specific sense of *barriera* while classified *corallo* according to the Semantic Type Natural Substance.

QA pair #7:
Q: *Qual è la capitale del Bhutan?*
A: *Lo scorso 24 ottobre, durante il quindicesimo round di colloqui a livello ministeriale,le due nazioni avevano sottoscritto nella capitale bhutanese Thimpu un accordo bilaterale.*

In this case, in IWN we find both the name of the country (Bhutan) and the adjective, kept together by the PERTAINS_TO relation, used in IWN to link a noun or an instance and a relational adjective: e.g. *musicale*/*musica* (musical/music), *presidenziale*/ *presidente* (presidential/president), etc. Among relational adjectives we also find ethnical relational adjectives, that we can link by means of this relation to the relative proper nouns (as in this case).
In SIMPLE-CLIPS, only the name of the country can be found, classified as a Geopolitical Entity.

QA pair #8:
Q: *Quale animale tuba?*
A: *Il musicista aveva ritratto, con pari esattezza visiva, il ruggito del leone, il cinguettio dell'usignolo e il tubare dei colombi.*.

Both in IWN and SIMPLE-CLIPS the *colombi* (pigeons) correctly belong to the Animal taxonomy, both at lexical level (exploiting the HAS_HYPERONYM and ISA relations and an intermediate node, *uccello*) and at ontological level.
In IWN, the semantics of the concept is quite well defined and suitable for the task at hand, since *colombo* is also the ROLE_ AGENT of the *tubare* (coo) verb.

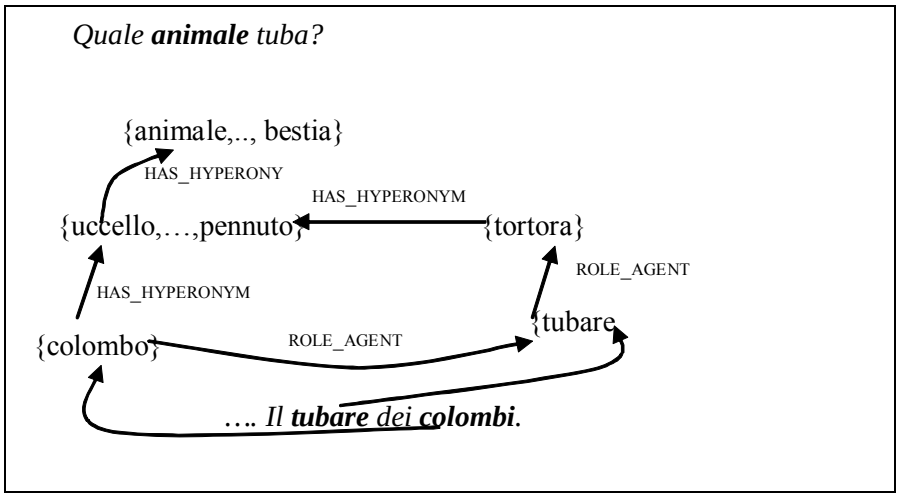


Fig. 32: semantics of *colombo* in IWN

QA pair #9:
Q: *Quale animale tuba?*
A: *I fidanzati tubavano sulle panchine, sussurrando dolci parole d'amore all'ombra degli alberi, giurandosi eterna e reciproca fedeltà.*

Here both the resources could be exploited to discard this paragraph as an invalid candidate answer. The subject of the verb, *fidanzato*, is in fact not indicated as an animal but as a human.

QA pair #10:
Q: *Quando e' stato stipulato il Trattato di Maastricht?*
A: *I commentatori hanno parlato a lungo della ratifica del Trattato di Maastricht avvenuta nell'autunno del 1992*.

In this case, the information in the IWN database seems to allow us to trace a path between the concepts of *ratifica* (ratification) and *stipulare* (to stipulate), a path that passes through the co-hyponym of ratifica, *stipula* (stipulation) that is a XPOS_NEAR_SYNONYM of the corresponding verb. It is important to notice that in this way *stipula* and *ratifica* as co-hyponyms should be considered mutually exclusive.

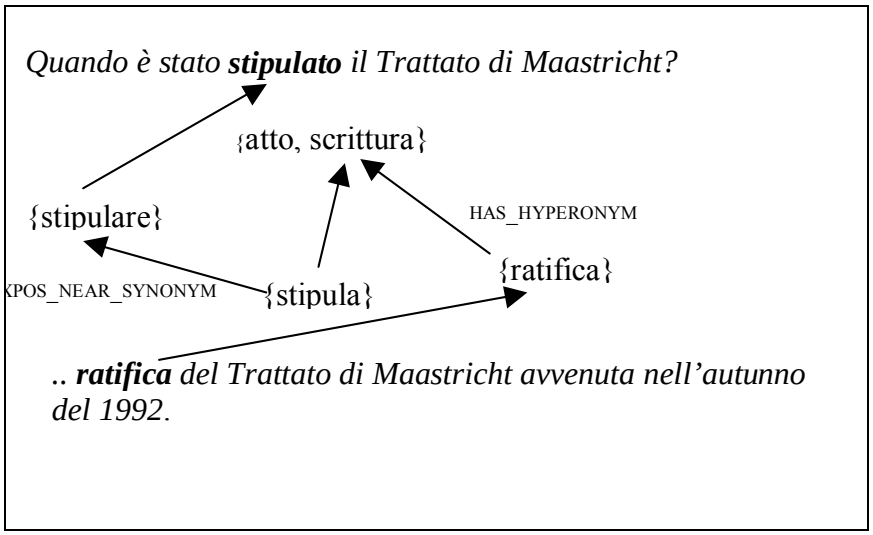


Fig. 33: connecting *stipulare* and *ratifica* in the IWN database

In SIMPLE-CLIPS *stipulare* and *ratifica* are kept together only by a common ISA relation to the very generic superordinate *agire* (to act), whereas they belong to different ontological types (the Relational Act and the Purpose Act).

QA pair #11:
Q: *Quanti membri della scorta sono morti nell'attentato al giudice Falcone?*

A: *la strage di Capaci, dove morirono il giudice Giovanni Falcone, la sua compagna Francesca Morvillo e tre degli agenti di scorta..*

The lexical mismatch in the QA pair that was so effortlessly handled by humans is represented in this way in the tow LRs:

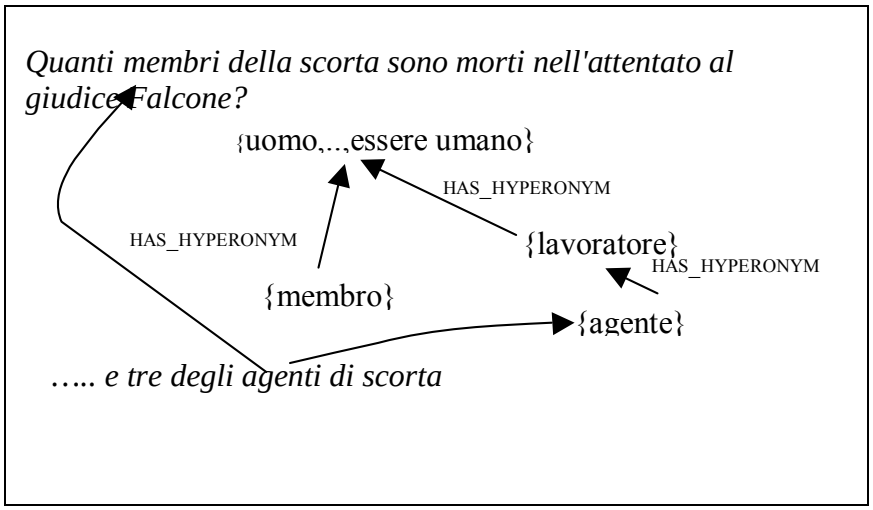


Fig. 34: connecting path between *membro* and *agente* in IWN

In SIMPLE-CLIPS, the two concepts (partly) share the same Ontological Types:

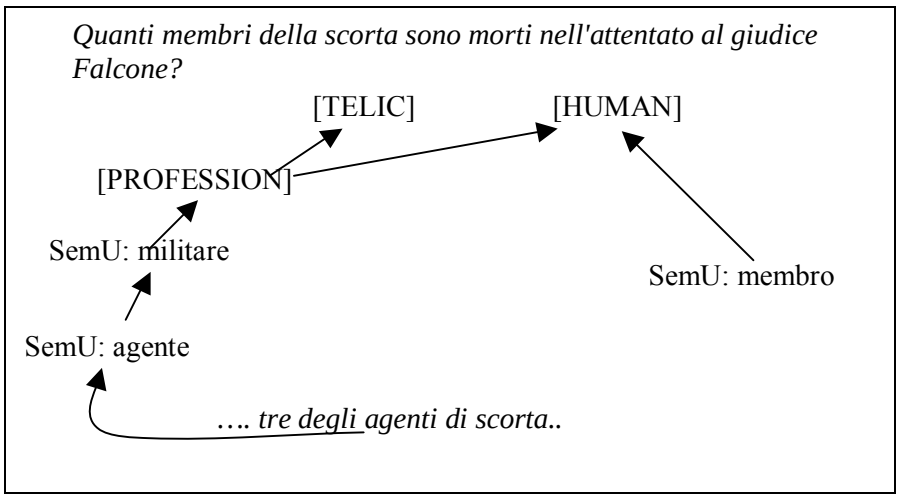


Fig. 35: shared ontological types in SIMPLE-CLIPS

QA pair #12:
Q: *Quanti anni di prigionia ha subito John Mendel?*

A: *John Mendel ha compiuto oggi una visita carica di dolorosi ricordi nel penitenziario di Robben Island dove egli subì 19 dei 27 anni di carcere.*

In this case, a path can be traced in IWN between the semantically close concepts of *prigionia* (detention) and *carcere* (prison).

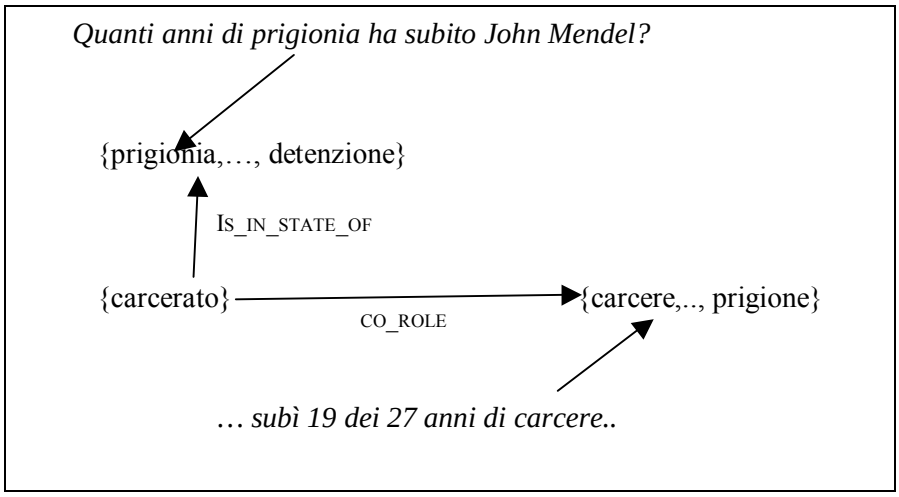


Fig. 36: portion of the IWN db dedicated to the prigionia and carcere concepts

The CO_ROLE elations were defined in EWN and also used in IWN to encode links between 1st order entities which have a role in the same situation: e.g., *pianista* (pianist) and *pianoforte* (piano) have a role in the situation referred to by *suonare il pianoforte* (to play the piano).

In SIMPLE-CLIPS no path was founded between the two SemUs.

QA pair #13:

Q: *Di quale nazionalità erano le petroliere che hanno causato la catastrofe ecologica vicino a Trinidad e Tobago nel 1979?*

A: *Al largo di Trinidad e Tobago (Mar dei Caraibi), entrano in collisione le navi "Atlantic Express" e "Aegean Captain", ambedue battenti bandiera liberiana.*

In IWN, both the adjective *liberiano* and the multiword *battere bandiera* are not present. In SIMPLE-CLIPS, the adjective (classified as Social Property) is present but no connection could be found in order to interpret it as a kind of nationality. The multiword is not present.

QA pair #14:

Q: *Quali esseri viventi sono in grado di assorbire l'anidride carbonica?*
A: *Secondo i meteorologi, i coralli sono in grado di assorbire CO2 e altri gas responsabili dell'incremento della temperatura del pianeta..*

In both the resources the chemical symbol corresponding to *anitride carbonica* is not present. This type of information is traditionally considered encyclopaedic and is usually not encoded in lexical resources.

QA pair #15:
Q: *Come vengono chiamati i piloti suicidi giapponesi?*
A: *Nella battaglia di Okinawa morirono più di mille piloti kamikaze che si gettarono sulle posizioni nemiche con gli aerei imbottiti di esplosivo e muniti della benzina sufficiente solo per il viaggio di andata..*

In both the LRs the semantics provided for the concept of kamikaze is not enough to operate efficaciously on this QA pair.
In IWN, the kamikaze is a hyponym of *aviatore* (aviator) that is in turn a hyponym of *pilota* (pilot).
In SIMPLE-CLIPS the kamikaze is only an Agent_of_Temporary_Activity.
The reference to the suicide nature of the kamikaze is present only in the definitions provided by the two LRs and therefore not exploitable by a system.

QA pair #16:
Q: *Come vengono chiamati i piloti suicidi giapponesi?*
A: *Il kamikaze giapponese, tanto bravo e veloce quanto sprovveduto, rompendo dopo pochi minuti il motore Yamaha della sua vettura ha inondato tutta la pista d'olio in maniera tale che si sarebbe potuta preparare un'insalata.*

Only in IWN a specific synset is available for this figurative sense of the word kamikaze, and only the different target of the hyperonymy relation can be exploited to distinguish the two senses.

QA pair #17:
Q: *Quando e' stato stipulato il Trattato di Maastricht?*
A: *La conclusione del Trattato di Maastricht è del 1991, anno ricco di avvenimenti importanti per l'Europa intera.*

In IWN, we found more than one synset for the word *conclusione*. In particular the first two senses are very close to each other and cannot easily be distinguished even by a human being. In order to trace a path between *conclusione* and *stipulare*, however, we choose the second sense of the word, obtaining the following path:

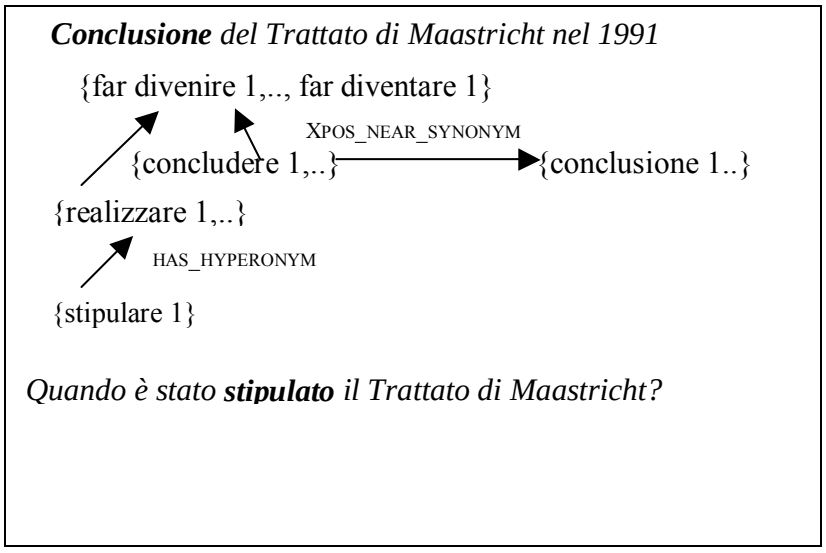


Fig. 37: connecting *stipulare* and *conclusione* in IWN

In SIMPLE-CLIPS, no path can be established between the two concepts stipulare and conclusione, that are also classified according to two different Semantic Types (respectively Relational Act and Causal Aspect)

QA pair #18:
Q: *Qual è il quotidiano nepalese più letto*?
A: *Fatturato complessivo in lieve calo per la stampa nepalese nel 1993 mentre il Corriere del Nepal si conferma al primo posto nella classifica dei quotidiani nazionali.*

Nothing in the two language resources seems able to help us to support the necessary inference that would allow us to recognize the fact that being in first place in the top ten means to be the most widely read. This information would be traditionally classified as world knowledge and in this sense not housed in a lexicon.

QA pair #19:
Q: *Chi si è addormentato durante il discorso di inaugurazione dell'anno giudiziario?*
A: *durante il discorso di inaugurazione dell'anno giudiziario il presidente del senato stava russando, con evidente imbarazzo del resto della platea.*

All the participants who took part in the questionnaire were able to effortlessly infer that *snoring* implies *to be asleep*. This same inference can be supported exploiting the IWN relation of the type IS_A_SUBEVENT_OF which links the *russare* and the *dormire* synsets.
On the contrary, nothing in SIMPLE-CLIPS helps us to establish this same connection.

QA pair #20:
Q: *In che giorno è stato ucciso Aldo Moro?*

A: *Aldo Moro è morto il 9 maggio 1978, tre mesi dopo il suo sequestro ad opera delle Brigate Rosse.*

We already demonstrated (cf. QA pair #2) that the required connections can be found in the LRs.

QA pair #21:
Q: *Dove risiede Gifrat?*
A: *Immagini inconsuete scuotono la coscienza di Israele e pongono domande difficili, domande che, sicuramente, a casa sua, a Gaza, Gifrat si pone per converso.*

We already demonstrated (cf. QA pair #3) that the required connections can be found in the IWN database. In this case, moreover, a suitable semantic path can also be found in the SIMPLE-CLIPS database:

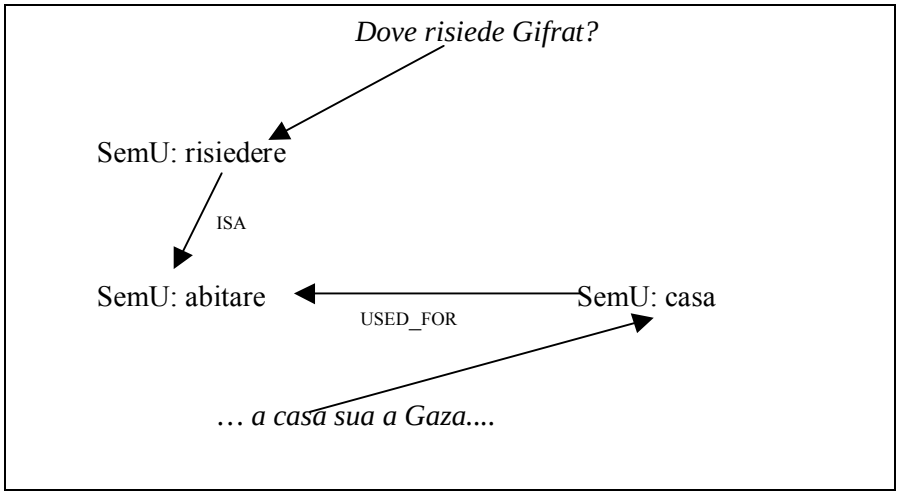


Fig. 38: connecting *risiedere* and *casa* in IWN

QA pair #22:
Q: Cosa può causare il tumore all'intestino?
A: *Studi recenti dimostrano come gli OGM causino il cancro all'intestino.*

In both the language resources a similarity (even if to a different degree) can be established between *tumore* (tumor) and *cancro* (cancer). While in IWN the two words belong to different synsets connected by means of a NEAR_SYNONYM relation, in SIMPLE *cancro* and *tumore* are indicated as synonyms.

Some results can already be discussed: potentially, the linguistic models of both lexicons seem able to provide some support to inference (always, with the exception of cases for which world knowledge is involved, like in case of QA pair#17). Sometimes, however, the needed link is missing because it was not encoded at all or because it was encoded in an alternative way with respect to the way it would have been useful for the specific task at hand. On 22 QA pairs, some form of connection between question and answer was found 13 times when using ItalWordNet and 7 times when using SIMPLE-CLIPS.

In the following paragraphs, we verify whether these connections can be exploited within an actual QA prototype. The fact that sometimes semantic paths potentially useful for the individuation of the answers can be detected does not guarantee that they are actually exploitable by an application. As a matter of fact, the mere presence of a path does not mean that that path is logically valid and computable. In 2.5.2.4 we describe a methodology presented in (Harabagiu and Moldovan, 1998 and Harabagiu and Moldovan, 2000) to automatically find semantic paths through semantic relations able to drive the matching between question and answer. Harabagiu and Moldovan's approach is important because it shows very clearly how not all the possible paths are bearers of meaning but only those that are composed by logically valid chains of relations.

4

In this chapter, we firstly introduce the construction schema of a QA prototype for Italian language. We analyse the baseline of its performance, obtained without lexico-semantic feedback. Then, we show at what extent the results can be improved by using information stored in language resources.

4.1 What we have learned so far

We will now try to sum up the "lessons" we have learned from the previous chapters. First of all, we have learned that existing QA systems successfully exploit a certain amount of information available in lexico-semantic language resources. We have also learned that usually only hyperonymy and synonymy are successfully exploited in QA and this can be read as a signal of the difficulties that arise when there is an attempt to exploit other types of relations. Nevertheless, we have also shown that sometimes the inferences and connections that humans so effortlessly perform when they identify an answer to a given question seem supportable by different types of relations encoded in the lexicons. Other times, on the contrary, lexicons seem not able to constitute a support to inference. We know that only one system exploits a wide variety of relation types and we have described the methodology adopted in the construction of the so-called inferential chains (cf. 2.5.2.4).

In order to verify the actual contribution of language resources and to analyse the nature of the difficulties that emerge during their exploitation, we will try to plug ItalWordNet and SIMPLE-CLIPS in a real Question Answer prototype for Italian. The next part of the chapter is thus dedicated to the preparation of what we can call the "experimental environment", i.e. the QA prototype. The construction of such an application is not something that involves only the access to the synsets and to the SemUs in the LRs. On the contrary, a QA application is the result of many different implementative choices concerning a variety of problems, ranging from the syntactic analysis of the question, to the creation of the query, to the integration of a Search Engine and the definition of strategies for the extraction of the answer. When we will introduce the overall architecture of the system, we will describe all these issues, still devoting more attention to the modules of the prototype where information encoded in LRs is much exploited. A detailed description of the diverse strategies adopted in building the prototype can be found in (Bertagna *et al*., 2005).

4.2 The Testbed

The first step in studying QA strategies for languages other than English is the creation of a benchmark of questions. When this research began, this benchmark for Italian was missing[27] altogether and the simple strategy of extraction of interrogatives from a large Italian corpus[28] was for the most a failure: the forms extracted are not the factual Wh-questions we are interested in but rather rhetorical and Y/N questions[29]:

Ripetevo qualche volta fra me con la sua voce gutturale e cortese: «Hai il papà?* Ma tu ce l'hai il papà?» Infine, smisi d'averne paura. Ma feci di lui - BO1989 Mai devi domandarmi C26.436.p.157 .4*

Come uno mette il piede fuori calpesta il prato, o scompiglia il ghiaietto. Vede quel piccolo ontano laggiú?* È tanto che vorrei andarlo a vedere da vicino, ma sono mortificato dalle pedate che resterebbero sull'erba. - BO1985 Atlante occidentale C5.73.p.52 .11*

energia, pura luce, pura immaginazione? Non vede come le cose ormai cominciano ad essere non-cose?* Come non chiedono piú movimenti del corpo ma sentimenti? Non piú gesti ma intelligenza, e percezione? Non - BO1985 Atlante occidentale C6.39.p.68 .12*

Some spontaneous factual *Wh*-questions (about 200) were extracted from web sites dedicated to on-line quiz. We decided not to use questions extracted from on-line FAQs, as the topics and the type of lexicon were too specific and domain-dependent. The major part of the reference corpus was built by translating into Italian the 499 questions of TREC-10 (2001). The original English questions of this wide test set are based on search logs donated by Micorsoft and Ask Jeeves. This first part of the reference corpus was used to study the most common *Wh* interrogative forms for Italian and the strategies to automatically analyse them syntactically. Moreover, the TREC-10 questions were carefully studied to understand how they might be classified according to a taxonomy of expected answer types.

In the meantime, we had the opportunity of using the Italian question collections of the CLEF 2003 and 2004 QA tracks as benchmark for the system. This opportunity was fundamental since the questions were accompanied by a large reference corpus of Italian newspaper articles where the answers to the questions can be found.

[27] No Search Engine log files containing Italian questions were available.
[28] A part of the PAROLE corpus of about 20 millions of words (cf. Marinelli et al., 2003).
[29] The contexts have been extracted using the DBT (cf. Picchi, 1991).

4.3 The two prototypes

4.3.1 Text meaning representation

Before introducing the system schema, we would like to discuss about what is the final text representation we want the system to achieve. It is a very important yet difficult task because we know that more than one possibility exists and that from it derive all the other choices (what are the resources we have to plug in the overall architecture, how the contribution to each analysis step can be merged in a unified representation, what are the syntactic and semantic clues that can be of help in the application flow etc.). Moreover, the system has to build not one but two text representations (of the question and of the candidate answers) that have to be mapped onto each other when the system has to "answer the question". Given the first question of the CLEF2004 test set:

In quale anno venne conferito il premio Nobel a Thomas Mann?

the following are a set of information that represents its meaning:

- that the question expects as answer a *specific year*
- that in some year the Nobel Prize was conferred to Thoman Mann
- that *Thomas Mann* and *premio Nobel* are phrases corresponding to proper noun;
- that *Thomas Mann* is first name and surname of a human being;
- that premio Nobel is an award
- that the award was won in the past

At the same time, given the candidate answer:

Davos (GR), 12 ago (ats) Si e' chiuso venerdi' il simposio di Davos che per cinque giorni ha visto riuniti nella cittadina grigionese 600 lettori dello scrittore tedesco Thomas Mann, premio Nobel della letteratura nel 1929.

The following are the set of information that represents its meaning:

- that we are dealing with a fragment of newspaper article
- that Davos is a name of location
- that 12 ago is a temporal expression
- that ats is the name of the press agency that produced the article
- that the expression Davos (GR), 12 ago (ats) is the time and place the article has been written
- that a symposium was held in Davos
- that the symposium ended last Friday
- that the duration of the symposium was five days

- that the symposium is used metaphorically as the subject of the verb vedere to represent the fact that in the occasion of that symposium 600 readers of Thomas Mann gathered.
- that Thomas Mann is first name and surname of a human being;
- that Thomas Mann is a writer
- that Thomas Mann is German
- that Thomas Mann was awarded with the Nobel Prize for literature in 1929.

Obviously, this set of information is just a "frozen snapshot" of the overall meaning of the texts of question and answer: more information may be added in subsequent and more granular representations, like for example that a symposium is a social gathering, that five days is an expression of time corresponding to 120 hours, that Thomas Mann, being a living entity of type animal, breaths, that probably the writer was awarded for a book he wrote, that this book, being the writer German, is probably written in German, etc. The introduction of (Bertuccelli Papi, 2000) provides a nice example of the way the meaning of a text is representable by making emerge one-by-one new levels of information, in a lacuna-filling and potential endless effort that tries to disclose the meanings implicitly present in the text. In that introduction to *implicitness*, the author shows the progressive emerging of new, hidden and subsequent meanings from a newspaper article talking about the Eurotunnel: the article is re-edited three times, everytime showing new particulars driven by inference. However, the author recognizes that:

> *[The n.d.r] undeniably more explicit [..] version of our text does not, however, extend to the point of uncovering the whole amount of meanings implicitly communicated by the text itself [..]*
> *(Bertuccelli papi, 2000)*

Meaning is not something that can be determined in a discrete way but rather a continuum. Nevertheless, we think that the lists of properties of question and answer we proposed above would provide a good basis for the processing of a QA application. The system has, after the text representations are provided by the analysis modules, to match the two representations, focussing on the particular portions of properties that adhere to the informative requirements of the question. In our case, for example, the system will have to understand that the expected answer is a year and it will have to verify that the occurrence of 1929 in the candidate answer refers to Thoman Mann awarded with the Nobel Prize.

The prototype is planned in such a way to show how the various information types that enrich, one-by-one, the text representation, can be exploited to individuate and extract the answer. The system thus constitutes the experimental environment for this research and it is organized following the classic three-module architecture consisting of the question analysis, the search engine and the answer extraction modules (cf. 2.4). In order to better explain the impact of the lexico-semantic feedback in this type of application, we organize this chapter into two sections: in the first one, we describe a first application, where no lexico-semantic information is exploited. We provide the results of this application on the testbde provided by the CLEF-2004 organizers and we consider these results as a baseline of the performance of the system. For this reason, we call this first version of the prototype the "baseline prototype". In the second part, we present a

second version of the application, where the three modules are (alternately) enriched with information available in the IWN and SIMPLE-CLIPS databases. The idea is to show what benefits are derived by this information in terms of performance improvements. We call this second version of the prototype the "lexically enhanced prototype". The results of the two versions will be presented and discussed.

4.3.2 The "baseline prototype"

The three-module architecture of the "baseline prototype" can be briefly described in this way:

- in the first module, an analysis of the question is performed in order to extract the information that will be of use in the QA stream, i.e.:
    - i) the list of the question keywords that will be used in the IR module,
    - ii) the Question Stem,
    - iii) the dependency representation of the question that will be compared against the dependency representation of the candidate answer,
    - iv) the Answer Type, i.e. the restricted set of "expected answer Types" that can be directly derived from the question stems Chi (Who), Dove (Where), Quanto (How much) and Quando (When).

- The second module consists of a document indexing and retrieval sub-system that receives in input the keywords of the query and provides in output a list of paragraphs matching the query .

- The last module is where all the information collected during the first phase of question analysis is exploited. A system of filters rules out candidate paragraphs not satisfying a certain set of constraints. In the "baseline version" of the prototype, a module exploiting the dependency structure of the question and of the candidate answer has been implemented, together with the exploitation of named entity types that can be individuated by means of simple pattern matching rules.

Fig. 39 represents the logic architecture we have in mind for the "baseline prototype"[30]: we can observe how the various modules of analysis interact with each other, in particular how the output of the Question Analysis module (an XML file where all the information from the morphological analysis, the syntactic parsers, the stemmer etc. are gathered and homogeneously represented) becomes the input for the Search Engine and how the output of the search engine becomes the input of the phase of answer selection and extraction. We designed this first layer of the application just as we think it should be, i.e. taking into

[30] The "work flow" was different from the "logical" information flow between the modules of the architecture. As a matter of fact, the first module available was the central one, i.e. the IR module, represented by the IXE Paragraph Engine (Attardi and Cisternino, 2001). We preferred to build the rest of the application by integrating this core module with all the other components of analysis, adding what was missing step by step in order to incrementally improve the partial system results.

consideration also modules of analysis that we do not have at our disposal but that we consider really fundamental. In the figure, we represented those modules by using the broken line: they are the module for Word Sense Disambiguation, for multiword recognition and the Named Entity recognizer. Their functioning will be simulated during our experiments. Another module would be highly useful: the one that, evaluated the output of the system, is able to decide about the correctness and pertinence of the result in order to give rise to alternative strategies involving query expansion.

In the next paragraph we describe in detail the various modules of analysis reported in Fig. 39, and introduce the exploited tools and resources. We also point out, for each step, in which "areas" lexico-semantic information would have been useful.

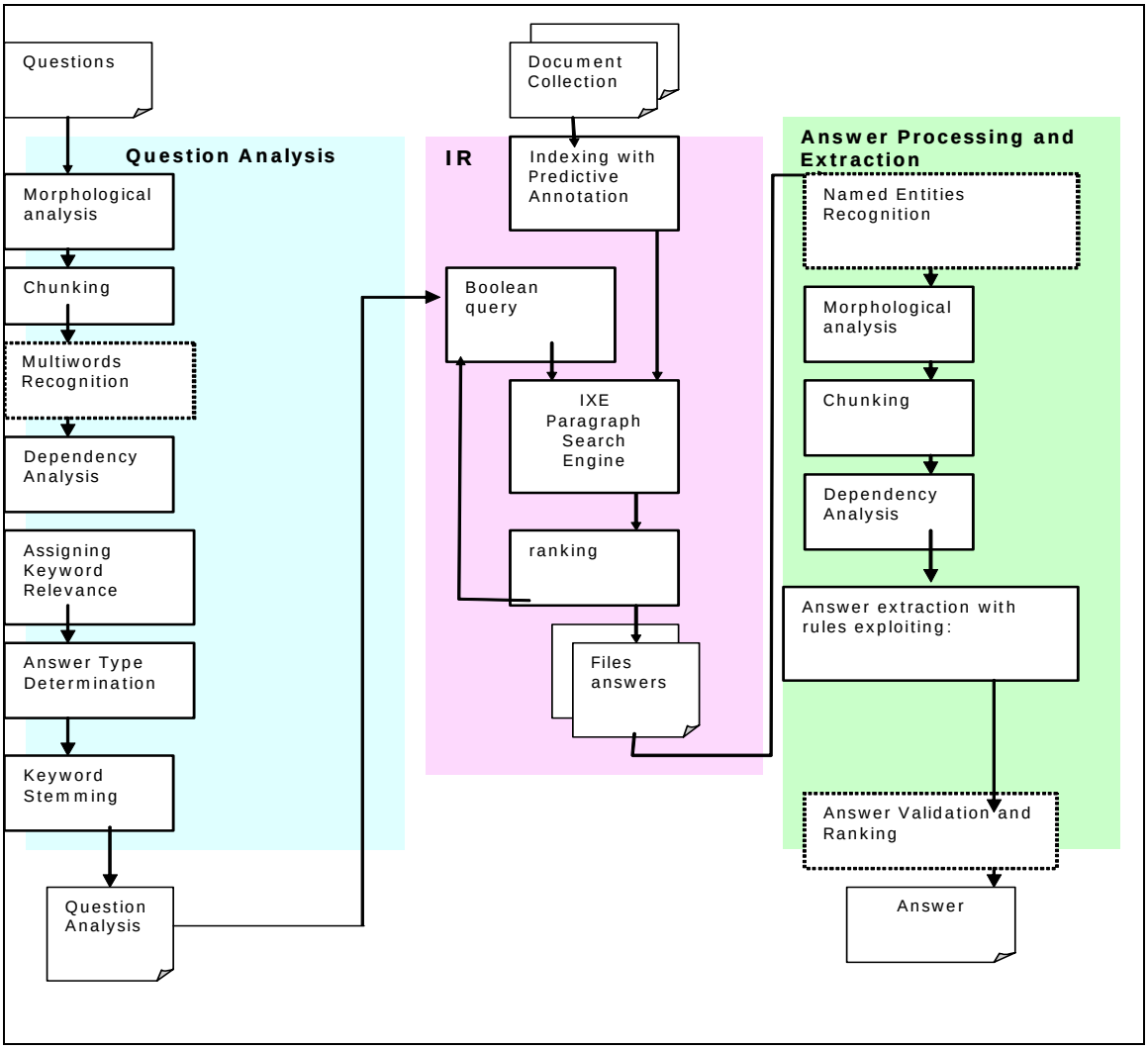


Fig. 39: Logical architecture of the "Baseline Prototype"

4.3.2.1 THE QUESTION ANALYSIS MODULE IN THE BASELINE PROTOTYPE

The following description provides an overview of the module that performs a multi-layered analysis of the question:

- First, a sequence of steps leads to the linguistic representation of the question: each word of the question is isolated, morphologically analysed and associated to one or more lemmas. Then a two-stage (chunking and dependency) syntactic analysis is performed, allowing the system to: i) segment the question into syntactically organized text units, ii) perform POS-tagging of the words in the question, iii) identify grammatical functions;
- The system applies a set of rules in order to assign to each word in the question a specific weight in term of its relevance as a keyword of the query;
- The system extracts the Question Stem (the interrogative element usually introducing the sentence) from the question.
- The Answer Type (i.e. the expected answer type) is individuated by merely relying on the Question Stem type;
- A stemmer is used on some of the keywords of the query.

The following paragraphs will describe in more details each of these steps.

4.3.2.1.1 Linguistic Analysis

First of all, the question goes through a chain of tools for the analysis of the Italian language developed at ILC-CNR (Bartolini *et al.*, 2002). The analysis chain includes:

i) Morphological analyser
ii) Chunker
iii) Dependency analyser

The morphological analysis is performed by Magic (Battista and Pirrelli, 1999). For each word form of the question, Magic produces all its possible lemmas together with their morpho-syntactic features. Magic also recognizes the capitalization of the word, a small set of basic multi-word expressions and analyses verbs containing clitic pronouns.

The chunker, CHUNK-IT (Lenci *et al.*, 2001), first performs the morpho-syntactic disambiguation of the question and then segments it into an unstructured sequence of syntactically organized text units (the *chunks*). We will see how even this initial, flat syntactic representation can be exploited to extract the Question Stem, that is crucial for the task of question classification on the basis of the type of expected answer (i.e. what the user is looking for with his/her question).

The chunked file is the input of IDEAL (Italian DEpendency AnaLyzer) that generates a representation of the sentence using binary, asymmetric relations (modifier, object, subject, complement etc.) between a head and a dependent, based on the FAME annotation schema (Lenci *et al*., 2000). The success of a QA application highly depends on the quality of the parser output and it is very important to efficiently parse interrogative forms and extract the syntactic relations that allow the system to recognize information such as direct object, subject etc. that have such an importance in the semantic interpretation of the sentence. Part of the activities of the current research was dedicated to the creation of a specific set of rules for parsing Wh-quesitons (starting from the analysis of a corpus of Italian interrogative forms).

The paragraphs returned by the Search Engine and candidate to be identified as answers will be subjected to these same linguistic analysis and tools.

We can say that the morphosyntactic and syntactic analysis is the key for an initial semantic interpretation of the question, aimed at deriving the expected answer type when the stem is evocative and the system does not have to semantically analyse the answer type term.

4.3.2.2 THE ANSWER TYPE TAXONOMY IN THE "BASELINE PROTOTYPE"

The types of expected answer are organized in a hierarchical structure that we call Answer Type Taxonomy (ATTax).

In order to understand what selection of nodes could be used to represent the variety of the possible expected answers, we have analysed about 500 questions of the testbde. We identified 42 different types of expected answer but the number can vary greatly since the classification can be more or less granular. Among the various identified Answer Types (ATs) we find for example ANIMAL, HUMAN, DEFINITION, COLOUR and many others.

Clusters of lexical-syntactic patterns compose the Answer Type Taxonomy. The patterns are typical of specific types of question and are organized in a taxonomic way. They are conceived to map different syntactic realizations into a same conceptual representation.

Some ATs can be determined via pattern matching on the question stem that allows us to get closer to the expected answer type and to the text portion that is likely to contain the answer. Some Question Stems, for example Quando (When) and Dove (Where), reveal which kind of answer we can expect to receive and a set of simple rules was encoded in order to allow the system to establish univocal correspondence between them and specific ATs. The following table shows some correspondence that can be established between stems and Answer Types.

Chi (Who) ➔HUMAN
Quando (When) ➔ DATE
Dove (Where) ➔ LOCATION
Perché (Why) ➔ REASON
Quanto (How Much) ➔ QUANTITY
Come (How) ➔ EXPLANATION

This correspondenc represents an over-simplification: for example, it is not true that all the questions introduced by the stem *Dove* have a location as the expected answer: CLEFquestion#118, *Da dove viene estratto l'acido salicilico?*, expects an answer regarding a substance or a concrete material and not a location. In the same way, the hypothetic question *Dove Dante parla di Francesca e Paolo?* asks about a literary work and not about a geographical location. Nevertheless, the very simple correspondence table is the only thing we can do by exploiting pattern-matching rules. Probably, even semantic language resources will not be enough to help the system to correctly derive the ATs of these questions, since we would need sources of what we can call *world-knowledge* information.

Moreover, in order to discover other ATs, the system detects some simple common patterns[31] which involve the first chunks of the questions; this is true, for example, for questions where the interrogative adverb *Quanto* (How much) is followed by the verb *pesare|durare|costare|misurare..* or by the sequence copula + adjective *alto|pesante|lungo|profondo...* These patterns give rise to some ATs such as WEIGHT, HEIGHT, COST, and LENGTH.

Another set of more specific pattern matching rules was written to allow the system to recognize some of the so-called DEFINITION QUESTIONS, i.e. the questions of the type *Chi è Silvio Berlusconi?* (Who is Silvio Berlusconi?), *Cosa è il Mossad?* (What is the Mossad?) or *Cosa è il diabete* (*What is diabetes?*). In the "baseline prototype" this type of question is identified by simply looking in the question for patterns of the type:

[(*Che*)+*cosa*] + copula + (Proper Name|Noun)
Chi + copula + Proper Name

In these cases, a specific Answer Type, DEFINITION, is assigned to the question.

The ATTaxonomy is, in this first release of the prototype, just a one-dimensional structure constituted by conceptually equivalent syntactic patterns with some lexical constraints. When we introduce the enhanced prototype, we will see that a semantic layer of information will be added to the Taxonomy, making the lexical elements in the patterns the starting point for the navigation of the word meanings present in language resources.

An important issue is how many ad hoc rules can be considered appropriate for the baseline system. Given the aim of the dissertation and of the research, we believe they should not exceed the number of links between the ATTaxonomy and the LR of the enhanced release: if the same effort in defining the constraints at lexical level in the two Taxonomies does not produce a significant improvement in the enhanced prototype, then it means that all the semantic information we add to the taxonomy is not really useful.

The following picture shows how the ATTaxonomy can be determined by exploiting only simple pattern matching rules without recurring to semantic information stored in LRs.

[31] The pattern matching is not directly performed on the text but on the output of the syntactic analysis.

TOP

FACT

TIME

Pattern: QS= Quando
Pattern: QS=quale|che, ATT=”periodo”

DATE

Pattern: QS=quale|che, ATT=”giorno”
Pattern: QS=quale|che, ATT=”anno”
Pattern: QS=quale|che, ATT=”mese”

QUANTITY

Pattern: QS= Quanto[adv|adj]
Pattern: QS=quale|che, ATT=”quantita”

DISTANCE

Pattern: Quanto+copula+ ATT=”alto|lungo|distante”
Pattern: Quale|che+copula+ ATT=”altezza|lunghezza|distanza”
Pattern: Quanto+ATT=”distare”

WEIGHT

Pattern: Quanto+copula+ ATT=”pesante”
Pattern: Quale|che+copula+ ATT=”peso”
Pattern: Quando+ATT=”pesare”

COST

Pattern: Quanto+copula+ ATT=”costoso”
Pattern: Quale|che+copula+ ATT=”costo”
Pattern: Quando+ATT=”costare”

SPEED

Pattern: Quanto+copula+ ATT=”veloce”
Pattern: Quale|che+copula+ ATT=”velocità”

HUMAN

Pattern: QS=Chi

LOCATION

Pattern: QS=Dove

COUNTRY

Pattern: Stem=Quale, ATT=”paese”

CITY

Pattern: Stem=Quale, ATT=”città”

REGION

Pattern: Stem=Quale, ATT=”regione”

CONTINENT

Pattern: Stem=Quale, ATT=”continente”

DESCR

REASON

Pattern: QS=Perché,
Pattern. QS=Quale, ATT=”ragione|motivo|causa”

DEFINITION

Pattern: QS=”chi” + copula + NP
Pattern: QS=”che cosa” + copula + N

DIFFERENCE

Pattern: QS: quale, ATT=”differenza”

EXPLANATION

Pattern: QS=Come

Fig. 40: The Answer Type Taxonomy in the Baseline Prototipe

This is the easiest way to recover these types of Answer Types. Following this simple method, the baseline prototype was able to recognize the Answer Type for 63% of the questions. A different strategy, exploiting the synonyms and the hyponyms of the ItalWordNet hierarchies will be presented, similar to the one presented in (Paşca, 2003) (cf. 2.5.2.1.2). We will compare the results of both approaches, in order to understand whether a "light" approach based on pattern matching techniques is enough to reach good results.

4.3.2.2.1 A hybrid taxonomy

Even if this Answer Type Taxonomy is still quite "poor" and "basic", it is already possible to realize how it is somehow hybrid in an ontological sense. As a matter of fact, we can see that it comprises two main types of ATs, i.e. the ones referring to:

- questions whose answers can be classified according to an ontology of types. In the case of the ATs Human or Location, for example, the expected answer can be classified respectively according to the types Human and Location. This means that questions like *Chi ha scritto la Divina Commedia?*, and *Chi ha scoperto l'America?* expect an answer regarding a human name (*Dante Alighieri* and *Cristoforo Colombo*), while *Francia* and *Atlantico*, as answer for *Dove si trova Parigi?* and *Dove è naufragato il Titanic?,* can be classified as name of locations[32]. The answers we expect from this type of question are usually a single entity, often represented by a Named Entity. These types of expected answer are grouped under the common top node FACT.
- questions whose answers consist of definitions or explanations. This is true, for example, for definition questions or for questions introduced by the stem *perché* and *come* (that ask for explanations, instructions, procedures etc.). Usually these types of questions require long, explanatory answers and are also inherently ambiguous since the kind and amount of explanation required is dependent on the user's information need. These types of expected answer are grouped under the common top node DESCR.

If we consider the "type of answer" as the discriminating factor in the distinction between definition/explanation and factual questions, we obtain that the difference between these two types of answer is not well-defined, as we can see by considering the definition question "*Che cosa è un colibrì?*": a fully informative answer to this question would be a definition (something like "*A tiny bird which moves its wings very quickly*") but also a shorter answer like "bird" or "animal" can be considered valid and correct. So, there is the possibility that the answer to a definition question corresponds to a single lexical item, a lexical item that is in a very particular relationship with the object of the question, i.e. it is its hyperonym. Nevertheless, there is a deep difference between i) asking about a definition of something, ii) asking about exact, factual

[32] Obviously, only the types of questions introduced by the "right" question stems are handled by the baseline prototype that is not able to derive the expected answer type of questions equivalent to the provided examples (*Quale poeta ha scritto la Divina Commedia?*, *Quale navigatore ha scoperto l'America?* and *In quale paese si trova Parigi?* and *In quale oceano è naufragato il Titanic*).

information that is missing in the cognitive description of a given known event or state and iii) asking about explanations and reasons behind the facts.

However, even if a kind of "ontological" difference between types of answer can be recognized, nothing prevents us from collecting them in the same taxonomy. We must not forget that the first use of such taxonomy is the possibility of "triggering" different strategies when the system recognizes what type of answer we can expect from a particular question.

The 22 ATs presented in Fig. 40 show what can be derived by solely recurring to the stem-based rules of the baseline prototype or by matching recurrent lexico-syntactic patterns. As already mentioned, only a limited number of specific patterns introduced by *Che* and *Quale* can be analysed and recognized since these two stems, being interrogative adjectives, do not provide any clues about the semantic category of the expected answer. In these cases, to obtain the expected answer type the system should analyse the noun modified by *Che* and *Quale* (the Answer Type Term) in order to derive the Answer Type (see paragraph 4.3.3.4) and this will be done in the "enhanced prototype". However, The Answer Type Term and the Question Stem are also derived in the baseline prototype (by recurring to the syntactic analysis of the question) since they are exploited in the module for the selection of the keyword (see paragraph 4.3.3.3).

4.3.2.3 THE PROBLEM OF KEYWORD RELEVANCE

We already mentioned (cf. 2.5.2.2) what factors were identified by (Paşca, 2003) as important for the selection of question terms as keywords: semantic salience, redundancy and degree of term variation. We think that semantic salience and redundancy are the two sides of the same coin, since what it semantically salient is not redundant for definition and vice versa. So, the problem of keyword selection is two-dimensional.

The following example (question#65 of the CLEF-2004 test set) is interesting because it allows us to observe the nature of the "keyword selection issue":

Al di sopra di quale area geografica è stato osservato il fenomeno noto come "buco dell'ozono"? (Over what geographic area has the phenomenon known as "ozone hole" been observed?)

This gives rise to many possible paraphrases with identical meaning of that same question:

1. *Al di sopra di quale area geografica è il "buco dell'ozono"?* (Over what geographic area is the "ozone hole"?)
2. *Al di sopra di quale area geografica è posizionato il fenomeno noto come "buco dell'ozono"?* (Over what geographic area is located the phenomenon known as "ozone hole"?)
3. *Al di sopra di quale area geografica si trova il fenomeno noto come "buco dell'ozono"?* (Over what geographic area is located the phenomenon known as "ozone hole"?)
4. *etc.*

In what follows, we give the two "answering" paragraphs returned by the Search Engine :

Sydney, 20 gen (ats/ansa) Il 'buco nell'ozono' sopra l'Antartide, che lascia passare i raggi ultravioletti cancerogeni, miete sempre piu' vittime in Australia che e' particolarmente esposta al fenomeno.

Il buco nell'ozono sull'Antartico si assottiglia tra ottobre e novembre e ogni anno si perde oltre il 60 per cento dell'ozono in una zona di 15-20 chilometri sopra l'Antartico.</answer>

Looking at the question and its paraphrases on one side and at the candidate answers on the other, we see that very few keywords "survive" in the passage from question to answer.

We would like to be able to isolate those lexical items that we will not plausibly find in the answer because they may be expressed with semantically close lexemes or even eliminated. (the case of *osservare* and its substitutes in 2, 3 and its elimination in 1). In other words, we may want to identify those words that are not informative but are in some way redundant for the essential meaning of the question (it is clear that also "fenomeno noto" can be dropped without important effects). This is the problem of semantic salience.

Other terms, on the contrary, are absolutely essential to the general meaning of question and answer, but may be expressed in different way: In the case of question#44, *Chi è l'inventore del televisore?* (Who is the inventor of television?), if we send to the Search Engine the expression *inventore* AND *televisore* we will not get the answer regarding *televisione*. In these cases, synonyms, hyponyms, hyperonyms, etc. are used in the answer instead of the original question keyword (this is the problem of term variation). The optimum in this case is to propose a list of alternative lexical items to the IR module and perform query expansion. Even if query expansion is usually done by exploiting lists of synonyms, the QA pairs of the questionnaire show that the types of lexical mismatches are varied and the exploitation of synonymy may not be enough. The ultimate goal is to collect a wider set of documents by sending to the IR module an alternative list of lexical items. The privileged measure would be *recall*.

As regards the second situation described, on the contrary, we want to avoid the recall of non-pertinent documents, i.e. documents where the semantically void terms of the question may appear. It is also important to identify redundant terms in order to avoid submitting them to query expansion (it would not make any sense to expand the adjective *noto*, known, with its synonymic expressions *conosciuto* or *famoso*). In this sense, we can say that the privileged measure would be *precision*. The problem, as we will see, is understanding which criteria allow the system to detect non-relevant items in the question.

The union of the two aims and strategies should give rise to the optimisation of the final result. In the following paragraphs we will describe the strategies adopted in the baseline prototype. We will introduce the more advanced modules of the enhanced prototype in par. 4.3.3.

4.3.2.3.1 Keyword selection in the baseline prototype: aiming at the essential.

The selection of the keywords for the query is a very important but difficult task. Since we do not have access to LRs in this level of the prototype, we cannot perform query expansion, nor can we try to detect distinctions of a lexico-semantic nature. Nevertheless, we can play with the many possible combinations of keywords in the composition of the Boolean query. The basic idea is *aiming at the essential*,

i.e. trying to isolate those terms in the questions that are really important. In the first question of the collection (*In quale anno venne conferito il premio Nobel a Thomas Mann?*[33]), we would like to submit a vector to the search engine containing *at least* the words: *Nobel*, *Thomas*, *Mann*. It is unlikely that we would find the word *anno* (year) in the expected paragraph (in its place we would more probably find the year we are looking for). Moreover, the noun *premio* is not indispensable to indicate *premio Nobel* (*In quale anno venne conferito il Nobel a Thomas Mann?*) while the word *conferito* can easily be substituted by a synonym (like *assegnato*, assigned) or by *vincere* (win) if in the answer *Thomas Mann* is indicated as the person who won the Nobel prize[34]. There seems to be a sort of "persistence scale" where the degree of lexical variation goes from a maximum (in the case of adverbs) to a minimum (in the case of Proper Nouns). The scale can be roughly represented in this way:

adverbs > Verbs > (Nouns, Adjectives) >Proper Nouns

Intuitively, we could think that abstract entities, like verbs, adjectives and adverbs, are the most subjected to phenomena of lexical variation. If so, it would be useful to assign a low relevance score to adjectives and abstract nouns in the question.

As for as adjectives are concerned, it is very unlikely that we will find a qualitative adjective in the question; on the contrary, very often the adjectives are important to precisely select the answer and to refer to the attribute the answer should have in order to fulfil the informative needs of the questioner exactly. When the question asks about the *first president of the United States* (question#30), what the questioner wants is simply the name of the first in history and not the second or third. In the same way, when asking about a *fundamental ingredient in Japanese cuisine* (question#52), we want to know the name of some Japanese ingredients and not something used in German food. For this reason, we decided to assign a high score to adjectives (the same used for nouns). But if it is true that adjectives used in questions submitted to a QA system are usually very salient under the semantic point of view, it is also true that they are highly variant, because they are often substituted by semantically equivalent expressions (for example *giapponese* by *del Giappone*, *alto* by *altezza* etc.) In the enhanced prototype we will try to improve our strategy concerning adjectives, by expanding the query with correspondent concepts (4.3.4).

Furthermore, the strategy consisting in assigning a lower relevance score to abstract nouns (information that could be retrieved by the ontological classification in LRs) does not provide the expected results. We analysed the 400 questions of the 2003 and 2004 editions of the CLEF campaign and, differently from what was expected, the analysis seems to disprove the initial assumption: only in very few cases would the abstract/concrete distinction have played a role in an effective selection of the keyword. Term variation and semantic salience do not seem to have anything to do with abstract/concrete opposition (at least for this specific task). It is true that sometimes the abstract noun of the question is substituted in the answer by a

[33] *What year was Thomas Mann awarded the Nobel Prize?*

[34] *Also this first example shows that, in order to deal with this task, the access to various types of information would be required. We should be able to access not only morphosyntactic and syntactic information (for the identification of the PoS and of the ATT) but also lexico-semantic information (synonyms or other variant of the keyword).*

synonym or a close term (as happens for CLEF2003-question#99, *Quanti membri dell'equipaggio sono morti nel disastro del sottomarino "Emeraude"?*, where the noun *disastro*, disaster, is substituted in the answer by the noun *incidente*, accident), but it does not happen more frequently than in the case of concrete nouns. The same can be said regarding the semantic salience issue: there are cases of abstract nouns that should be dropped from the query but not more than concrete nouns. Some nouns should be discarded from the query because they are kind of stopwords (for example *nome* in *qual è il nome…?*), others are used in adposition to better specify and define a Proper Name (*affezione* in *Dammi un sintomo con cui si presenta l'affezione da virus Ebola*, *serie* in *Chi interpretava James Bond nei primi episodi della serie 007?*) but not in a way substantially different from what happens with concrete nouns (see gruppo in *A quale età Michael Jackson ha cominciato a cantare nel gruppo dei "Jackson Five"?*). 81 of the 400 questions of the test sets comprehend abstract nouns, but only in very few cases they should not have been sent to the Search Engine. On the contrary, there are plenty of cases where the abstract noun has a prominent role in the question and cannot be discarded in place of a noun with a concrete referent. For example in CLEF2004-question#186 *(Chi è il ministro della sanità francese?*[35]) the abstract noun *sanità* (health) has a fundamental informative role in the question (and in this case the adjective is too highly discriminating). The same thing happens for CLEF2004-question#?? (*Quando è stata approvata la convenzione sui diritti del bambino?*), where the concrete *bambino* is not more important than *convenzione* and *diritto*[36]. When introducing the enhanced prototype, we will see that a different criterium of semantic nature can be adopted, the one concering the generality/specificity issue.

In the "baseline prototype", in order to deal with the majority of cases, we adopted a general rule on the basis of the different Parts Of Speech and of the syntactic function of the word in the question (by exploiting the output of the chugger and of the dependency parser). The basic idea is to send to the Search Engine different combinations of keywords in subsequent loops: at the beginning, the majority of the terms in the question (with the exception of stopwords) are sent to IXE. Then, loop-by-loop, the (supposedly) less important keywords are dropped or composed in OR and at the end only the (supposedly) very important keywords are used in the query.

To each morphological word an attribute "relevance" is assigned which is set to the minimal value (0) if the word belongs to a list of stopwords and to the maximum value (10) if the word is a number, has a capital letter (*Quante esecuzioni capitali ci sono state negli Stati Uniti nel 1993?*) or is in inverted commas (*Che cosa ha influenzato l'"effetto Tequila"?*). The Part of Speech of the remaining words is analysed and an intermediate value (7) is assigned to the relevance of nouns and adjectves while a smaller value (5) is assigned to verbs and adverbs (the minimum value, 0, is assigned to auxiliary or modal verbs).

Other rules apply to more specific yet frequent cases, for example assigning the minimum value to the relevance of the verb *chiamare* in question#121 (*Come si chiama la moglie di Kurt Cobain?*[37]) or of the

[35] Who is the Frech Ministry of Health?
[36] It should be noted that both questions contain sort of multiword expressions (*ministro della sanità* and *convenzione sui diritti del bambino*) that, consequently, should be treated without any decomposition.
[37] *What is the name of Kurt Cobain's wife?*

verb *trovarsi* in question#134 (*Dove si trova l'arcipelago delle Svalbard?*[38]). Also questions introduced by various imperative verbs like *nomina*, *dammi*, *dimmi* etc. are dealt with.

Other more subtle distinctions may be introduced but are not essential for the current discussion[39].

All the nouns that are "answer type terms" in questions introduced by the interrogative adjectives *Quale* and *Che* and by the pronoun *Quale* (the word *anno* in the question *In quale anno venne conferito il premio Nobel a Thomas Mann?* and the word *professione* in the question *Qual è la professione di James Bond?*) received a low score (2), as did their modifiers. This because it is plausible that at their place we will find the answer we are looking for. At a first glance, a different strategy seems to be more suitable for questions introduced by *Quale* with a pronominal function. As a matter of fact, many questions of this type seem to require that their ATTs be sent to the IR module. This is the case of question#26: *Quale è la capitale della Russia?*, whose ATT *capitale* is often present in the text of the paragraph so it would be good to use it as keyword of the query:

(AGZ.951015.0049) *"…Il sequestro di un autobus di turisti sudcoreani conclusosi la scorsa notte a Mosca con la morte del rapitore e la liberazione degli ostaggi e' il primo del genere nella capitale russa e il primo che coinvolge cittadini stranieri.."*

Since at a first glance the difference seems to be of "semantic" nature, we will thoroughly analyse this topic when we introduce the keyword selection module of the enhanced prototype.

4.3.2.4 STEMMING

The Porter stemmer (Porter, 1980) for Italian[40] was used on all the keywords with relevance smaller than the maximum value (so in general only Proper Nouns and keywords in inverted commas were not stemmed). The use of a stemmer was preferred because it seemed simpler and more straightforward than the automatic generation of morphological forms, but it has some important drawbacks (see paragraph 4.3.2.8.2 in the baseline results).

Stemming techniques are alternative approaches to morphological expansion and papers (Bilotti *et al*., 2004, Monz 2003a) are dedicated to assess which of the two approaches is the best (with different conclusions). However, stemming also has an interesting "spin-off" for semantics since it can expand the query not only to morphological variants (like in the typical case of different inflections of a verb) but also to lexical items that are semantically related to the original term. This is the case, for example, of the keyword *premio* (prize) of question#1, that, when stemmed, becomes "prem", thus enabling the retrieval of documents

[38] *Where is the Svalbard archipelago?*
39 for example, the first name is more optional than the surname in the retrieval of the paragraphs and this is the reason for the failure of retrieval for question#28 (Qual è il titolo del film di Stephen Frears con Glenn Close, John Malkovich e Michelle Pfeiffer?) where all the names with capital letters are submitted together (connected by AND) to the Search Engine while in the answer only the surname of John Malkovich is present.
[40] Available free at http://snowball.tartarus.org/italian/stemmer.html

containing related words such as the verb *premiare* and the adjective/past participle *premiato*[41]. Another example may be the keyword *amministratore* (administrator) in question#2 that, stemmed as "amministr", allows the search engine to also retrieve documents containing the verb *amministrare* (to administer) or the noun *amministrazione* (administration). *Premiare* and *premio, amministratore* and *amministrare* and *amministrazione* are semantically related and, in a lexical semantic resource, their connection may be represented by recurring to a relation of the *role* type (and also, obviously, by derivational relations). If no such a resource is available, an interesting alternative could be constituted by the stemming operation. The most interesting "pro" for such an approach is the fact that it is really simple and "light" under a computational point of view. Obviously, however, it has strong limitations since the stemming option allows the retrieval of only the terms where the semantic connection is accompanied by the morphological derivation. Moreover, the possibility of expanding the query to related terms is restricted to words longer than the word in the query. In fact, if the question keyword is *giapponesi* (Japanese), the word *Giappone* (Japan) cannot be retrieved since it is shorter than the stemmed keyword "giappones*". We will see that this type of information can be derived from the relations encoded in LRs, which can be bi-directionally exploited.

4.3.2.5 QUESTION XML DATA STRUCTURE

In order to collect all the information derived from the various steps of question analysis, we recurred to an XML representation. Fig. 41 shows an exemplar question represented in our XML data Structure[42].

```
- <question clef_id="D IT IT 0008" q_id="q_8">
    Chi e' Shimon Peres?
  - <words>
    - <word cl="M1" relevance="0" value="chi" w_id="q_8w_70">
        <morph forma="chi" m_id="q_8w_70m_1" others="!,_,_,!,!,!,pos" pos="pron" value="chi" />
        <morph forma="chi" m_id="q_8w_70m_2" others="!,m,p,!,!,!,pos" pos="nn" value="cha" />
      </word>
    - <word relevance="0" value="e'" w_id="q_8w_71">
        <morph forma="e'" m_id="q_8w_71m_1" others="3,!,s,pres,ind,!,!" pos="v_fin" value="essere" />
      </word>
      <word cl="M1" relevance="10" value="shimon" w_id="q_8w_72" />
      <word cl="M1" relevance="10" value="peres" w_id="q_8w_73" />
      <word punc="Y" relevance="0" value="?" w_id="q_8w_74" />
    </words>
  - <chunks>
      <chunk AGR="@FP@FS@MP@MS" CC="N_C" POTGOV="CHI#P@FP@FS@MP@MS" c_id="q_8c_1" />
      <chunk AGR="@S3" CC="FV_C" POTGOV="ESSERE#V@S3IP" c_id="q_8c_2" />
      <chunk AGR="@MS@FS@MP@FP" CC="N_C" POTGOV="SHIMON#SP@NN" c_id="q_8c_3" />
      <chunk AGR="@MS@FS@MP@FP" CC="N_C" POTGOV="PERES#SP@NN" c_id="q_8c_4" />
      <chunk CC="PUNC_C" PUNCTYPE="?#@" c_id="q_8c_5" />
    </chunks>
  - <relations>
      <relation dep="CHI[1]" head="ESSERE[2]" plaus="100" r_id="q_8r_1" role="PERSON" type="SUBJ" />
      <relation dep="SHIMON[3]" head="ESSERE[2]" plaus="100" r_id="q_8r_2" type="PRED" />
      <relation dep="PERES[4]" head="SHIMON[3]" plaus="100" r_id="q_8r_3" role="APPOS" type="MODIF" />
    </relations>
    <stem value="chi" />
    <question_focus value="ROLE" />
  </question>
```

Fig. 41:The Question XML Data Structure

[41] Many are the other words beginning with "prem-". Sometime, the cooccurrence of the keywords imposed by the Boolean query limits the possibility of recall of not pertinent terms but it does not happen always.

[42] It would be very useful in the future fully exploiting the *ids* of the various layers of linguistic representation in order to better represent the links between morphological forms, chunks and the heads/dependents of the functional analysis. This would facilitate the identification of the text portion containing the answer in the answer extraction module.

4.3.2.6 IR MODULE

We already talked (Chapter 2) of the importance of the presence of an effective retrieval subsystem in the overall QA architecture: if the IR module fails to find any relevant documents for a question, further processing steps to extract an answer will inevitably consequently fail.

The inner part of the system consists of a passage retrieval application built on a search engine developed at the Computer Science Department at the University of Pisa. The search engine, the same used in the PiQASso (Attardi *et al*., 2001) document indexing and retrieval sub-system, is based on IXE (Attardi and Cisternino, 2001), a high-performance C++ class library for building full-text search engines.

The search engine stores the full documents in compressed form and retrieves single paragraphs (in chapter 2 we learned that in QA this strategy is preferred to the full document indexing). However, full documents are also indexed and sentence boundary information is added to the index, to enable a wider search to nearby paragraphs. In fact in some cases all the relevant terms do not appear within a paragraph, but some may be present in nearby sentences. If the option to search in a wider context is chosen, those terms may still contribute to the retrieval and ranking of the paragraph. It very frequently happens that the answer takes more than a single paragraph. Nevertheless, in the two versions of our prototype, we preferred to restrict the search to the single paragraph, since we do not have the possibility of handling and treating fuller answers, for which, at least a module of anaphora resolution would be necessary.

4.3.2.6.1 Query formulation

The strategy followed to retrieve the candidate answers consists in the iteration of the Boolean query on the basis of the score "relevance" of each keyword and of the number of retrieved documents. In the first loop we send all the keywords to the Search Engine with relevance higher than 2 connected with the AND operator. If no paragraph is retrieved then the system performs the second loop, creating a query connecting all the keywords that have relevance higher than 7 with AND and with OR all the keywords with relevance 5. If no paragraphs are retrieved then the system performs the third loop. This consists in a query with all the keywords with relevance 10 in AND and the keywords with relevance 5 in OR. Again, if no paragraph is returned then the fourth and last iteration is performed with only the keywords with relevance 10.

The system also foresees a mechanism to restrict the proximity, in the case of queries that contain a sequence of first names and surnames (so the keywords *Thomas* and *Mann* of question#1 are searched for in the paragraphs without any other elements in between). This scheme has to be revised and inserted in the future in the more general strategy for handling poly-lexical units of the type name+surname, name+preposition+name (the *Mostro di Firenze* of question#48) etc.

4.3.2.6.2 Predictive Annotation Feature in IXE and the "Named Entity Recognizer issue"

A new release of the IXE Search Engine is under development at the Uni-Pi Computer Science Department: it allows queries constrained with information about the expected answer type, so for example in the case of question#3 (*Chi è l'amministratore delegato della FIAT?*) it is possible to submit a query of the type "amministratore delegato person:*" and retrieve only paragraphs containing the name of person. This technique is called *predictive annotation* (Prager *et al*., 2000) and consists in the identification of potential answers in texts by accordingly annotating and indexing them. This feature was not available at the time of the CLEF-2004 campaign and it is still under refinement also at the current stage of the research. It is based on the possibility of having at one's disposal a good Named Entity Recognizer able to tag, at indexing phase or during the extraction of the answer, the textual material using a set of common types. Named Entity Recognition is usually carried out by exploiting on-line *Gazetteers* (that have some drawbacks, like being static, i.e. intrinsically incomplete) or some form of feature learning (see results of the Named Entity Recognition task of the Message Understanding Conference at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/).

The availability of such a technology is of primary importance for the successful implementation of a QA system: as a matter of fact, even the best question analysis module is not useful if the system is not able to recognize the name of a person, an organization or a location. We will see (cf. 4.3.2.7.1) that many answers can be extracted by recurring to other methodologies. For example, our system mainly exploits syntactic dependency relations; the problem is that regularities in syntactic context are rare so it is not always easy to exploit rigid syntactic-based rules. The Named Entity Recognizer allows the implementation of more flexible rules, allowing, for example, the extraction of answers of the Location type in the case of questions of the corresponding type (generally, the method also foresees the contribution of other heuristics, like for example the consideration of the ranking of the paragraphs and of the mutual proximity of the various keywords). The importance of a support for NERecognition is also evident since for about 68% of the questions of the CLEF2004 test set the expected answer is a Named Entity.

Thus, the NERec should always be present in the QA pipeline, as a support for a predictive annotation approach or as semantic filter in the answer processing and extraction module. All the results we give in this dissertation are obtained by simulating the functioning of a Named Entity Recognizer able to detect instances of the type Person, Organization, Location, Date, Year, Date, Time, Money, Length, Weight, Speed. We want to remember, however, that the most advanced QA systems, such as FALCON (Harabagiu *et al*., 1999) can reckon on NE Recognizers able to work with dozens of Named Entity categories. In the prototype, only a small module was actually developed, by recurring to simple pattern matching on the text of the paragraphs. Moreover, the element "Named_entity" was created in the XML of the answer files.

4.3.2.7 ANSWER PROCESSING

The Search Engine returns a file for each query. The file returned follows a specific DTD having the paragraph as sub-element and the information about the match and the source document as attributes. No more than 40 paragraphs were saved in the answer files. The attribute "best_ranking" is also created at root element level, equivalent to the number of keywords actually submitted to IXE for the current query. For each paragraph, the system also calculates the value of the "ranking" attribute, consisting in the number of keywords of the query found in each single paragraph.

The meta-information representing the *coordinates* of the journalistic article (i.e. who wrote the article, where and when and for which news agency) are eliminated from the text in order to provide a *clean* input to the text analysis tools and are saved in a specific sub-element of type "MetaInfo". The paragraphs are then submitted to the morphological and syntactic analysers and the results are saved in specific elements.

4.3.2.7.1 Answer Detection and Extraction in the baseline prototype

Answer Detection is a very important module from our point of view: given an ordered set of paragraphs, the system has to establish which one is closest to the question. This task is very similar to the Textual Entailment problem, as defined in (Dagan and Glickman, 2004):

> *Textual entailment [..] is defined as a relationship between a coherent text T and a language expression, which is considered as a hypothesis, H. We say that T entails H (H is a consequent of T), denoted by T ⇒H, if the meaning of H, as interpreted in the context of T, can be inferred from the meaning of T.*

In our specific task, the H and T texts are our question and the paragraph returned by the Search Engine. (Pazienza *et al*., 2005) describes the possible cases that may occur:

1. T semantically subsumes H,
e.g. H= The cat eats the mouse and T= the cat devours the mouse;

2. T syntactically subsumes H
e.g. H= The cat eats the mouse and T= the cat eats the mouse in the garden

3. T directly implies H
e.g. H= The cat killed the mouse and T= the cat devours the mouse

In the first case, we see that the verb in H is the hyperonym of the verb in T while in the third case the verbs of H and T are connected by a "cause" relation. We realized that the the typology of connections can be wider, as the examples of the questionnaire show, with entailment further confused (both at the predicate and arguments level) by the use of synonyms and other lexical and semantic relations. Nevertheless, this short prospect gives a clear (even if simplified) idea of the situation we have to handle. We will describe the

exploitation of the semantic relations available in ItalWordNet and in SIMPLE-CLIPS when we introduce the enhanced prototype; a that moment, we will see how to handle the first and the third cases. For the moment, we have to focus our attention on the baseline prototype, that allows us to assess what (Dagan and Glickman, 2004) call a primary research interest, i.e. " '*how far' one can get by performing* [the required N.d.R.] *inference directly over lexical-syntactic representations, while avoiding semantic inference over explicit meaning-level representations*".

The second example deals with a case that can be resolved via pattern matching on syntactic dependency relations: H and T share the subject and the object, with a different specialization created by the presence of a complement of the verb in T. Syntactic pattern exploitation is thus one of the ways to match questions and answers (by focusing on the analysis of the relation which involves the question stem).

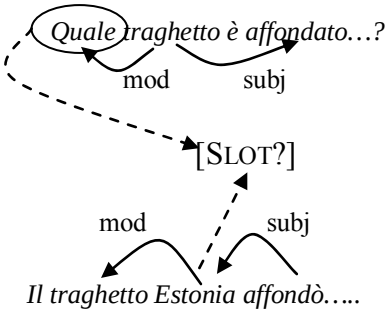


Fig. 42: dependency relations involving question stem

Differently from what happens in the case of the textual entailment task, however, the matching of a question and an answer is also heavily influenced by the restrictions on the type of expected answer (Fig. 43).

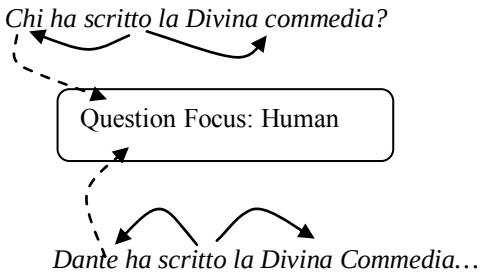


Fig. 43: matching dependency structures and restriction on the expected answer type

Thus, beyond the matching of dependency structures, other conditions are tested in order to individuate the answer among many candidates: i) named entities present in the paragraphs, ii) relative ranking of the paragraphs and iii) particular patterns in the answer text.
The rules are hierarchically organized and, when possible, are accompanied by a confidence score according to the degree of reliability of the provided answer. The adopted strategies will be described in more detail in the next paragraphs.

4.3.2.7.2 Dependency relations

The dependency analysis of the question allows the system to search in the paragraphs for significant relations that can be interpreted as clues for answers.
The simplest strategy (the first one that the system applies) consists of searching among the many syntactic relations of the paragraphs in order to find the same links that involve the question stem. The question stem can be interpreted as a *void slot* that can be filled with the answer, thus all the relations that have the stem as a target are very important.

The adopted strategy consists of a three-step search: first a paragraph where all the the relations expressed in the question are present. In the case of the following question:

Quale presidente nordcoreano mori' all' eta' di 82 anni ?

the system looks in the paragraphs for:

mod ([slot?], presidente)
mod (presidente, nordcoreano)
subj (morire, presidente)
comp (morire, età)
comp (età, anni)
mod (anni, 82)

In the case of question:

Quante persone affondarono quando l'Estonia si ribaltò ed affondò?

The system looks for:

mod([slot?], persona, type=card)
subj(persona, affondare)
subord (affondare, quando)
subj (Estonia, ribaltare)
subj (Estonia, affondare)

These are very optimistic attempts that gave no results when applied to the entire CLEF2004 testbde.

The second search involves all the relations that involve the stem, the noun or verb it is connected to. Again, in the case of the question:

Quale presidente nordcoreano mori' all' eta' di 82 anni ?

the system looks in the paragraphs for:

mod ([slot?], presidente)
mod (presidente, nordcoreano)
subj (morire, presidente)

In the case of question: *Quante persone affondarono quando l'Estonia si ribaltò ed affondò?*

mod([slot?], persona, type=card)
subj(persona, affondare)

If even this second search does not match any paragraph, a last iteration is performed by looking only for the relations targeting the stem (in the case of the previous two questions, respectively the relations mod ([slot?], presidente) and mod([slot?], persona, type=card). Given the extreme variability of the way the information is represented in the texts, this last iteration is the most exploited. We have to remember, however, that the Search Engine helps the system to select a small subset of paragraphs that should already be quite close to the question text, thus the exploitation of these unique relations is often enough to individuate the possible answer.
Obviously, rules like these (based only on the stem and on the purely syntactic form of the question) are very rigid and it is not easy to be so lucky as to find an answer formulated as a declarative version of the question. Nevertheless, it is surely worth a try and, in the case of success, the answer is assigned with the high confidence scores, which go from the maximum (10) to a minimum (5).

More flexible rules, specific for some types of question, are the ones based on the assumption of a certain level of correspondence beween specific ATs and particularly frequent patterns of syntactic descriptions. In the case of the AT Human, for example, a successful strategy consists of looking for relations of coordination and of modification of type adposition. The baseline prototype, for example, answer to CLEF2004question#2 (*Chi è l'amministratore delegato della Fiat?*) by detecting the coordination present in the paragraph:

…Nel corso dell'assemblea dell'Ugaf, a cui ha partecipato anche l'amministratore delegato della Fiat, Cesare Romiti,…

In the case of AT Location, the system searches among the complements of the keyword introduced by the preposition di (of) or in (in). This is the case of CLEF2003question#111: *Dove si trova la moschea di Al Aqsa?*, and its answer (*Gerusalemme*) that can be extracted from the paragraph:

… il diritto di pregare senza alcuna limitazione nella moschea al Aqsa di Gerusalemme Est, terzo luogo santo per gli islamici.

An answer identified by recurring to expected patterns of syntactic relations is probably a right answer but syntactic regularities are quite rare and the rules depend too much on the quality of the parser output. The exploitation of dependency relations is very useful also because, when it can be applied, it allows the sytem to avoid the obstacle of the recognition of named entity classes. If the question asks about a ferry (*Quale traghetto affondò al largo dell'isola di Uto?*) and in the answer we find the relation between ferry and Estonia (*il traghetto Estonia affondò et.*.), the system does not have to disambiguate the type of named entity of Estonia.

4.3.2.7.3 Named Entities

When it is not possible to rely solely on syntactic clues to individuate the answer, it is of vital importance to have the possibility of exploiting the Named Entities corresponding to the Answer Type of the question.

4.3.2.7.4 Pattern matching on the text of the paragraph

In the case of *definition* questions, the baseline system follows a very simple strategy consisting in the extraction of the text between brackets that follows the keyword. Also for other types of questions, like in the case of the AT Location, an attempt is made based on the extraction of the text between brackets. In this way we are able to answer CLEF2004question#20: *Dove si trova il campo di sterminio di Auschwitz?*, from the paragraph

…Un corteo di un centinaio di persone – composto di monaci buddisti giapponesi, rappresentanti delle comunita' religiose di ebrei, cristiani e mormoni - e' partito oggi da Auschwitz (Polonia) diretto a Hiroshima…

4.3.2.7.5 Paragraph ranking

When no other ways to individuate the answer can be found, the system answers with the highest scored paragraph[43].

4.3.2.8 BASELINE RESULTS

The methods to evaluate a QA application are different from the ones used for normal IR systems. Usually, the provided measure is precision, determined by taking into consideration the number of correctly answered questions on the total number of questions. In presenting our results, we adopted the same items and classification used by the CLEF organizers to provide the results of the 2004 campaign. We thus remember here some of the criteria and methods used in the CLEF experience of that year (for a more detailed description cf. Magnini *et al*., 2004 and the CLEF 2004 Guidelines[44]).

First of all, only one answer per question is allowed. Answers can be classified in four ways:

1. Correct (right, R), when the answer is clear and responsive;
2. Inexact (X), when the quantity of provided information is more than essential,
3. Unsupported (U), when the answer string contains a correct answer but not supported by the document from which it was extracted,
4. Wrong (W), when the answer does not fulfil the informative needs expressed in the question.

Moreover, questions are classified as factoid (F) or definition (D) and results are accordingly presented.

Some attempts were made to automatically evaluate the performance of QA systems (Breck *et al*., 2000) but most of the time, like in our case, systems are manually evaluated, with laborious and time-consuming work.

In the following table we summarised the results of the baseline prototype45.

#Answer	#Right	#Wrong	#IneXact	Overal Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy	
							Precision	Recall
200	91	87	22	45.5	42.7	70	0.62	0.5

Table 10: baseline results

[43] In the case of two or more paragraphs having the same ranking score, the system simply provides the first paragraph as an answer.
[44] available at clef-qa.itc.it/2004/guidelines.html
[45] The results are slightly different from the ones obtained for the CLEF-2004 competition and presented in the Proceeding of the conference (Bertagna et al., 2005). As a matter of fact, for the CLEF competition some modules exploiting the semantic analysis were already implemented.

In order to more carefully analyse the baseline results, we provide other measures. First of all, we give the percentage of wrong answers for each type of question (Table 11); then, we try to evaluate the performance of the system before the extraction of the answer. This last measure allows the evaluation of the strategies of keyword selection adopted in the baseline system.

4.3.2.8.1 Answers and types of questions

In Table 11 we provide the final results organized for question stem. Where necessary, we divided the results for factual and definition questions.

Question Type	**# of questions in the test set**	**% Wrong**
Quanto (adv)	1	100
Quale (pronoun)	17	76.4
Come	12	58.3
Quanto (adj)	18	55.5
Quanto (pn)	9	55.5
Quale/ Che (adj)	43	46.5
Others (dimmi, dammi, nomina)	7	57
Dove	14	35.7
Chi	35	34.2
Cosa (DEF)	10	33.3
Come si chiama	6	33.3
Quando	14	21.4
(Che) Cosa (pn)	14	25

Table 11: answered questions classified according to question stem

The most evident thing is that, as we expected, the system is not able to respond to many of the questions introduced by the interrogative adjectives and pronouns *Quale* and *Che*. With the exception of the very frequent cases that can be resolved by recurring to ad-hoc rules, in those cases the system cannot extract the answer since it "does not know" what type of entity it has to look for in the paragraph. Examples of these questions are:

In quale genere musicale si distingue Michael Jackson?,

Qual è l'ingrediente base della cucina giapponese?,

Di quale nazionalità erano le petroliere che hanno causato la catastrophe ecologica vicino a Trinidad e Tobago nel 1979?

Quale è la professione di James Bond?

A quale età Michael Jackson ha cominciato a cantare nel gruppo dei Jackson Five?

Al di sopra di quale area geografica è stato osservato il fenomeno noto come "buco dell'ozono"?

Etc..

In the same way, the questions introduced by imperatives such as *nomina* (name), *dimmi* (tell me), *dammi* (give me), cannot be answered without recurring to sources of information that help the system to analyse the semantics of the ATT. The same is true for semantically similar types of question, i.e. the ones introduced by the patterns "*Qual è il nome*…" and "*Come si chiama*…".

nomina una compagnia petrolifera
dammi il nome di una persona accusata di pedofilia
dimmi il nome di una catena di fast food
Come si chiamano i piloti suicidi giapponesi?
Come si chiama la moglie di Kurt Cobain?
Come si chiama la casa discografica di Michael Jackson?
Come si chiama la compagnia di bandiera tedesca?

Etc.

Very often, however, as a result the system is at least able to provide the entire paragraph containing the answer: this happens when the system exploits the "extreme measure" consisting of providing as an answer the paragraph with the highest ranking (i.e. with the higher value of the attribute "best_ranking"). So, 13 questions introduced by the interrogative adjective *Quale* were evaluated inexact (not wrong) because they contain the correct answer but also other text. This is the case, for example, of CLEF-2004question#155: *Di quale squadra di calcio francese era presidente Bernard Tapie?*. IXE extracted the paragraph:

Nuovi momenti difficili per l'industriale francese Bernard Tapie, ex ministro delle aree urbane, deputato e presidente della squadra di calcio di Marsiglia, l'Olympique (OM).

The system was not able to identify the AT HUMAN GROUP but was however able to provide an answer to the question.

It is thus important to highlight that about 38% of the *Quale* questions can be answered without any support from Language Resources but only by means of a good mix of keywords. Obviously, the answer is longer than what is needed but it fulfils the informative needs of the potential user. Nevertheless, in the majority of cases, the simple evaluation of the paragraph ranking is not sufficient to identify an answer. CLEF-2004question#3: *Qual è la città sacra per gli Ebrei?* is an example of not answered questions. As a matter of fact, in that case, the heuristic based on the first paragraph with the highest score retrieved a paragraph that talks about Gerusalemme but without explicitly mentioning it:

<answer document="AGZ.940517.0135" match="(32,53)" ranking="3" ref="2">Israele, che ha occupato la parte araba della citta' nel 1967, ha proclamato nel 1980 l'intera citta' sua "eterna ed indivisibile capitale" in quanto piu' importante luogo sacro degli ebrei. .</answer>

The same happens for CLEF-2004question#70: *In quale città si trova la basilica di San Pietro?*, for which the system provided as answer the first of the 44 paragraphs, i.e.:

> <answer document="AGZ.950416.0044" match="(34,37)" ranking="1" ref="2">
> A causa della mattinata piovosa e del freddo, la messa papale del giorno di Pasqua e' stata spostata all'interno della Basilica di San Pietro, pur essendo stato predisposto fin da sabato l'altare papale sul sagrato antistante il tempio per la celebrazione sulla piazza, con l'ornamento di centomila fiori olandesi. .</answer>

Obviously, there is no scientific reason for the highest paragraph to contain the answer. It is simply that the question often asks about information that is salient to the combination of keywords of the question. This means that, if we are talking about the Nobel Prize to Thomas Mann, we likely find the year of the award in the pertinent paragraphs.

As we said, however, this is not always true and sometimes the information we are looking for is literally buried under tons of non-pertinent paragraphs. This frequently happens when we submit queries consisting of only one keyword to the IR module: in these cases, the query may be too underspecified and the IR module may return too many paragraphs. The problem is when the system does not have any chance of pinpointing the answer in such a bulk of information (information that it handles in a completely indistinct way: all the paragraphs are the same, the only distinctive attributes are the ranking and the proximity among the keywords). This is the case, for example, of CLEF2004-question#31, *Qual è la professione di James Bond?*. In that case, only the name *James Bond* was submitted to the Search Engine that retrieved more than one hundred paragraphs. There is no hierarchy among these paragraphs: the subset is completely opaque and indistinct, all the text fragments simply contains the string “James Bond”, with the same ranking and the same proximity. In these cases, the baseline prototype is not really able to detect which is/are the paragraph/s that contains the answer. We will try to exploit the hierarchical and ontological information available in IWN and Simple-CLIPS in order to introduce in the enhanced prototype some heuristics to distinguish which paragraphs contain or can contain the answer.

4.3.2.8.2 Precision and recall in the IR module

In general, however, the possibility of finding the answer (or also the short paragraph that contains the answer, like in the cases above) is feasible only if the system is provided with a reliable procedure of keyword selection that allows the retrieval of subsets of paragraphs where the answer is present. But the presence of the answer in the subset of paragraphs is not enough, it would also be better to reduce the number of paragraphs returned by the search engine: it is a good balance between precision and recall that determines the chances of success. The validation of the output of the Search Engine is very important because it gives us the possibility of understanding how well the system works before the answer extraction procedure. In order to assess this aspect of the problem, we analyse the results of the system at the level of Search Engine output. We see that in 21% of the times, the answer is not contained in the paragraph of the subset.

Sometimes this is due to the submission of not pertinent keywords to the Search Engine or to incorrect PoS assignments: CLEF2004-question#12 (*A quanto ammonta il numero dei profughi palestinesi che si sono rifugiati in Libano?*) is an example of a question where both the cases are present: the verb *ammontare* and the noun *numero* were sent to the Search Engine while the adjectival reading of the word *profugo* (refugee) was preferred, thus lowering the relevance score of a very salient keyword. The result is that only *numero* and *Libano* were sent to the Search Engine, which obviously did not return any useful paragraphs.

Furthermore, the adoption of stemming techniques has some negative effects: For example, question#127 (*Quale animale tuba?*[46]) was badly processed because the only keyword sent to the Search Engine was *tub** (the Answer Type Term *animale* was correctly omitted in the query vector). For this reason, the Search Engine retrieved a lot of non-pertinent paragraphs, such as paragraphs talking about ***tub****eri* (*tuber*) or ***tub****ercolosi* (*tubercolosis*). This would be avoided by using the morphological expansion in place of the stemmer, even if this would obviously not preventing the retrieval all the documents regarding the musical instrument *tuba*[47]. Moreover, the stemming, being a method to expand the query, can sometimes determine the loss of an important paragraph in the first positions of the ranking: in query 74, for example, the question "A quanto ammonta la popolazione degli USA?", the keyword *popolazione* was stemmed and transformed in *popol**: in this way, almost 190 paragraphs were extracted but the "right" one was well beyond the forty positions taken into consideration. Among the many returned paragraphs we found totally non-pertinent information, such as:

USA: popol*arità Madonna in calo, preoccupata la Warner.*

If the keyword had not been stemmed, the system would have been able to find the answer in the 30th ranked paragraph.

4.3.2.8.3 Short and Long questions

The analysis of the results shows that both long and short questions are difficult to treat. Long questions are "dangerous" because it is not easy to efficiently combine the various keywords in the query and not obtain results which are too fine-grained. On the contrary, short questions can be hard to treat because the result could be too large to be handled efficiently. One of the most difficult cases is represented by short questions in which a keyword with the highest relevance score is accompanied by a single keyword with a low or medium relevance. For example, CLEF2004-question#18 (*Che lingua si parla in Germania?*) is transformed into the query "parl* Germania". In the 100 paragraphs returned by the IR module there is no

[46] *What animal coos?*

[47] We didn't explore the possibility to discard non-pertinent paragraphs on the basis of the different PoS of the keywords in the answer and in the question (we didn't find any existing systems that adopt a similar strategy). Such a filter would not allow the retrieval of paragraphs containing terms conceptually relevant even if belonging to different PoSs.

trace of the language spoken in Germany and the answer can be found elsewhere, expressed without the verb *parlare*:

> *LASTAMPA94-014112 43815 (230, 232)* Il secondo volume - uscita prevista fine gennaio '95 - si occupera' dell'Italia; il terzo della Francia e della Spagna; il quarto dei Paesi di **lingua tedesca**: **Germania**, Austria e Svizzera; il quinto dei Paesi dell'Est; il sesto di Gran Bretagna, Scandinavia e Paesi Bassi.

This is also an example of cases for which the adopted strategies did not have the expected effect: as a matter of fact, in that case it would have been better to submit to the Search Engine also the answer type term *lingua*, to which we assign a very low score (2). The same consideration can be made for CLEF2004-question#17 (*A quale partito apparteneva Hitler?*): in that case, only the verb and the Proper Noun were sent to the IR module, which retrieved three paragraphs about *ideas and bones belonging to Hitler*, but not about *Hitler belonging to a party* (while in the answer we found *partito nazista*, with no mention to the verb *appartenere*). This is due to the decision to treat all the keywords with the same relevance score identically. The ATT is thus discarded because its relevance is lower than the one of other verbs and nouns in the question. An approach which could deal with all these cases would consist in creating a first query with the ATT and the most relevant keywords in the question, in order to allow the system to extract the possible answer by looking in the paragraphs for patterns of the type ATT+modifier. Moreover, the pattern Quale + Noun in the question suggests that the answer can be sought among the modifiers of the noun (the answer is in the "same place" as the interrogative element). Nevertheless, we see that, in these cases, no help would derive form the use of LRs (we will try to investigate, however, whether LRs could be exploited in the case of correct paragraph retrieval).

Other times, finding the reason for the system failure is quite complex. This is the case, for example, of CLEF-2004question#30 (*Chi fu il primo presidente degli Stati Uniti?*). With the correct query "primo presidente stati uniti" we get a number of results, no one correct, like for example:

> *Clinton sollecita negoziati di pace entro novembre quando – primo presidente degli Stati Uniti – compirà una visista nell'Ulster.*
>
> or
>
> *Bill Clinton è il primo presidente degli Stati Uniti che abbia mai fatto visita all'emirato*

The right answer (*George Washington*) is instead in a paragraph where the very relevant *Stati Uniti* is not present because it is spanned in neighbouring paragraphs:

> *AGZ.940217.0069 il primo presidente George Washington era un massone e un proprietario di schiavi"*

In order to deal with this type of situation the system should be equipped with more sophisticated modules of analysis (based on *anaphora resolution*) capable of discarding *Bill Clinton* as an answer and of correctly evaluating information spanned in neighbouring text fragments.

Sometimes, however, the query is correct, all the keywords are correctly weighted but, notwithstanding this, the returned paragraphs do not contain the sought answer. There can be two reasons: there is no answer in the text collection (this can happen, and in this case the system should respond NIL) or the words used in the answer are different from the keywords submitted with the query. The last is the reason for the failure on question#44 (*Chi è l'inventore del televisore?*[48]), where the paragraph containing the answer is not retrieved since it does not contain *televisore* but its synonym *televisione*.

I tre autori, giovani giornalisti della Stampa, lo hanno dedicato allo scozzese John L.Beird, l'inventore della televisione, senza rancore.

Furthermore, in these cases the baseline prototype cannot do anything to retrieve this paragraph because between the question processing phase and the Search Engine the system does not perform query expansion. It is up to us to demonstrate that Language Resources can make the difference.

4.3.3 The Enhanced prototype

This second part of the chapter represents the core of this research since it is dedicated to the description of the so-called "enhanced prototype", i.e. the system whose functionalities are enriched with lexico-semantic information. What we have tried to do is to support the system with the same type of information that proved its usefulness in the applications already existing for English and described in chapter 2. Moreover, we have tried to verify whether the whole range of semantic links and paths which emerged during the analysis of the questionnaire (cf. chapter 3) can be exploited to bridge the gap between the form of the question and of the paragraph containing the answer.

The overall prototype can be conceived as a layered architecture, where the lower layer of functionalities is represented by the baseline prototype and the upper layer constitutes the enhanced prototype. The experiments are carried out by alternatively exploiting the two lexicons. This for more clearly evaluate the actual contribution of each lexicons and for being able to isolate the problems that emerge from the exploitation of the lexico-semantic information available in the two language resources. It is also possible to suppose a contemporary use of ItalWordNet and SIMPLE_CLIPS, in order to exploit the points of strength of each lexicon (for example ItalWordNet could be use for its synonyms while SIMPLE-CLIPS for the information concerning the predicative representation). What it does not seem advantageous is the contemporary exploitation of the same information type.

Fig. 44 shows the final architecture of the enhanced system; a comparison with the architecture of Fig. 39 shows that the innovation is constituted by the new role played by LRs in various modules of the system but also by the feedback the application is able to provide to the two lexicons.

[48] *Who is the inventor of the television?*

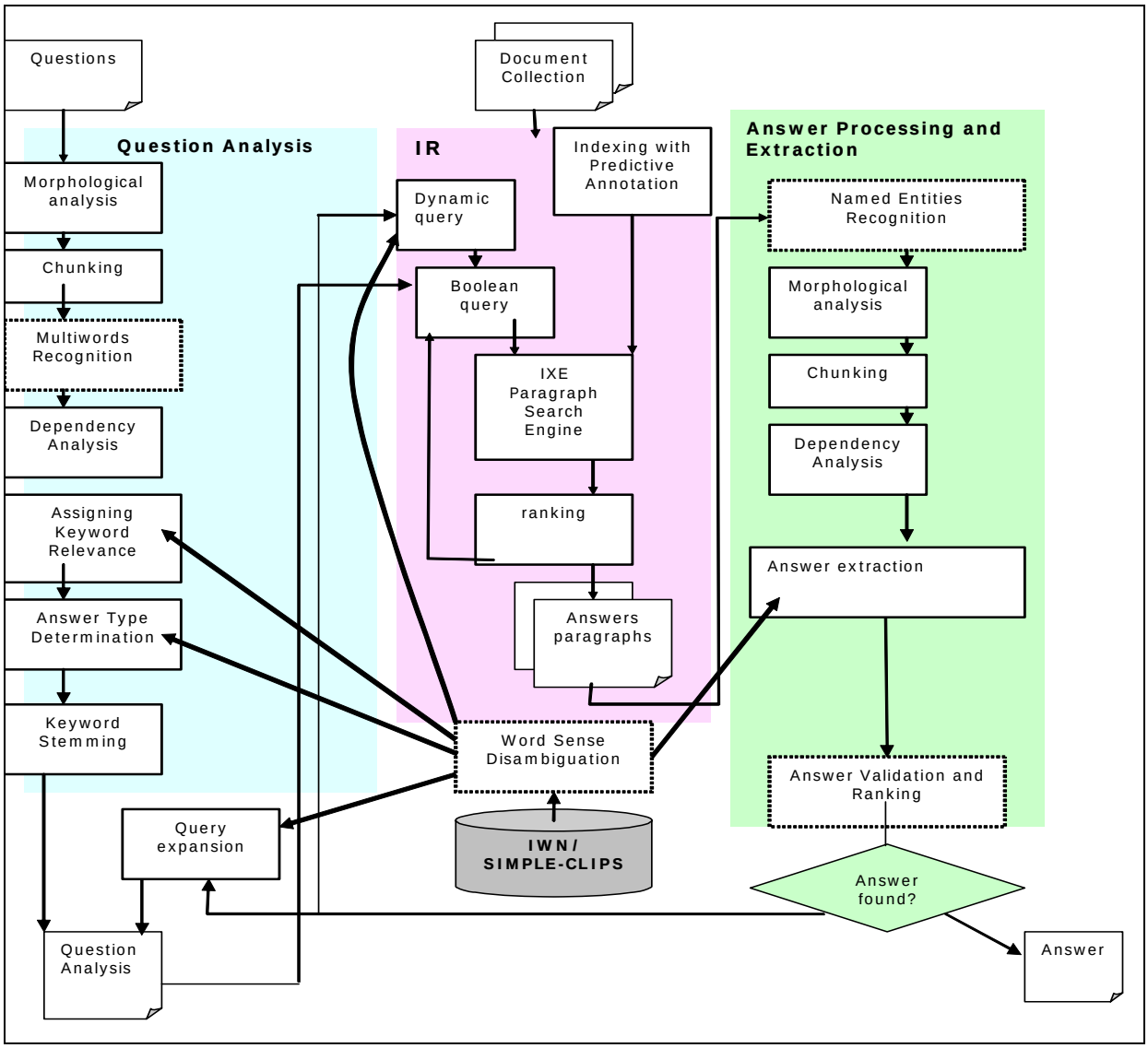


Fig. 44: overall architecture of the enhanced prototype

An alternative view is provided by next figure, where the connections between the data and processing flows and the static resources (both software and data) are represented.

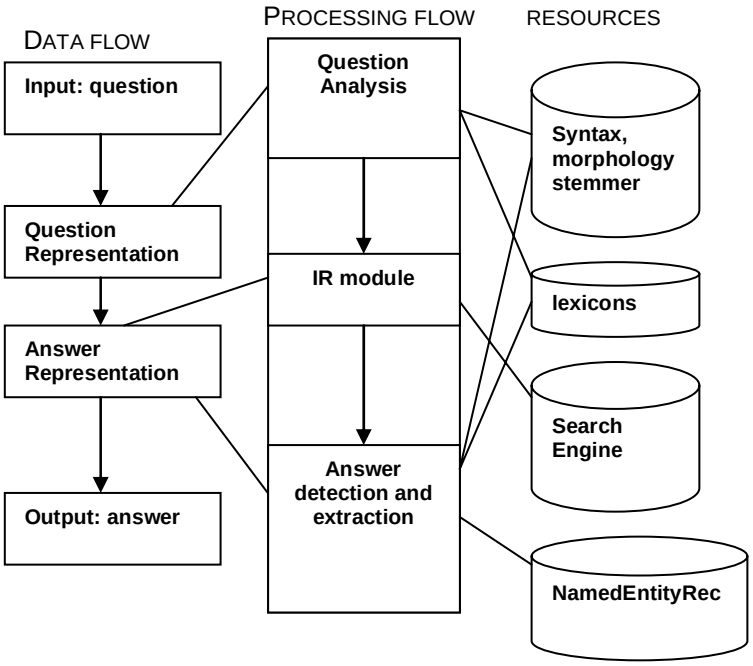


Fig. 45: data and processing flows and involved resources

4.3.3.1 A CLOSED MODEL THAT INTEGRATES DYNAMIC AND STATIC MODULES

We would like to realize a *closed* model that integrates language resources and procedural functioning of the application: in this model, not only the content of the lexical entry is exploited in the application but the application itself is able to dynamically enrich the lexical entry. A close-up of the integrated model we want to realize can be observed in Fig. 46.

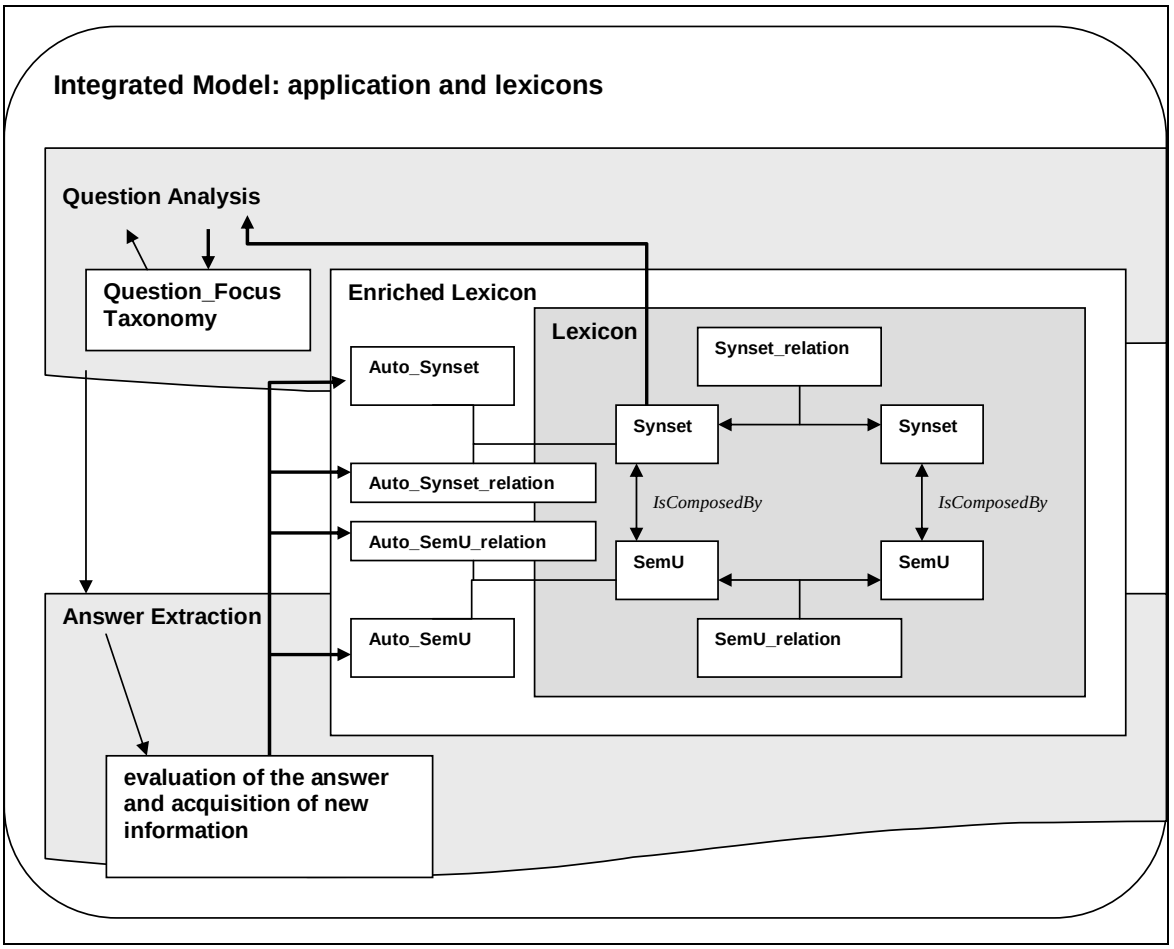


Fig. 46: integration of processing modules and static resources

Fig. 46 represents the lexicon as a two-layered architecture, consisting in a core lexicon (the original set of entries and relations manually or semi-automatically created) and in dynamically acquired information (that contributes to enrich the lexicon). The application incorporates the lexicon and exploits the overall available information (residing both in the core and in the external layer) in various moments of its three fundamental modules. In particular, the static content of semantic lexicons is used in the following modules of our system:

i) assignment of relevance to the keyword,
ii) determination of the Answer Type,
iii) query expansion,
iv) answer detection.

The output of the system (i.e. the answer) is automatically evaluated and, on the basis of the type of question and strategy adopted to extract the answer, it is added to the hierarchies of the lexicon as a new entry (in Fig. 46, we called these entries Auto_Synset and Auto_SemU). The entry can be already present in the lexicon but the system may also candidate different link between entries (what we called Auto_SemU_relation and Auto_Synset_relation).

In the following pages, we introduce all the innovations of the new version of the prototype. We will then analyse the results we get with these new functions in order to evaluate the impact of LRs on the whole architecture.

4.3.3.2 A PERVASIVE NECESSITY: WORD SENSE DISAMBIGUATION

A fundamental issue is that of Word Sense Disambiguation. Obviously, this is a pervasive problem or, as we can better say, it is "the" problem we have to face every time we want to access the content of LRs.

WSD is necessary in all the steps where LRs are involved. WSD is useful in the module for the determination of the Answer Type, if we do not want to make the system derive multiple ATs: for example, for question#155 (*Di quale squadra di calcio francese era presidente Bernard Tapie?*[49]) without any sort of WSD the system would identify, beyond the correct HUMAN GROUP, an incorrect AT INSTRUMENT, determined by the fact that the ATT *squadra* also has the meaning of *square*. Actually, we do not think it would be a very important limit for this specific task: the Information Retrieval phase should work as a kind of *implicit word sense disambiguator* since, in general, the co-occurrence of more than one keyword submitted to the Search Engine should determine the extraction of pertinent paragraphs excluding other readings (in this case, for example, no instruments can be found in the paragraph extracted: *Nuovi momenti difficili per l'industriale francese Bernard Tapie, ex ministro delle aree urbane, deputato e presidente della squadra di calcio di Marsiglia, l'Olympique…*[50]). However, this is not always true and the presence of very frequent types of occurrences in texts, like for example locations or human names, could determine the system failure in the case of erroneous derivations of AT HUMAN or LOCATION.

The situation is not different in all the other LR exploitation modules: for example, in the determination of keyword relevance, the system has to assess the specificity of the lexical item and in order to do so it has to individuate the right sense of the word. WSD is also very important in the creation of dynamic queries, which exploit the hyponyms of the ATT (see 4.3.3.4.6). Where the WSD is really indispensable is in the module for query expansion. In that case, the consequences of sending the wrong sense of the word to the IR module with all its semantic variants could be dramatic.

Not having the possibility of exploiting a WSD system based on complex features, we decided to rely on the assumption which claims that the individuation of the most frequent sense (the dominant sense in Kilgarriff, 2004) is enough to disambiguate a good percentage of occurrences. (Kilgarriff, 2004) reports on the work by Gale, Church and Yarowsky (1992), which identifies the so-called lower bound for the

[49] *Of which French football team was president Bernard Tapie?*
[50] *..Bernard Tapie, former minister for urban areas etc…*

performance of a WSD system as the score that a system achieves when it simply chooses the commonest sense (in their experiment they found an average score of 70%). (Kilgarriff, 2004) also defines the results of (Gale, Church and Yarowsky, 1992) as "a cloud sitting over WSD": if the results of such a mediocre type of system are so good, it becomes hard for an intelligent one to perform significantly better. It is basically true: the most frequent sense heuristic is the baseline for the evaluation of the performance of the systems that participated in the tasks of the Senseval campaigns. While in the case of the lexical sample task[51] of the Senseval-3 experiment (Mihalcea *et al*., 2004) the performance of most systems was higher that the baseline assessed between 55.2% and 64.5%[52] (with the best system performing at 72.9%-79.3%) [53], for the English all-words task (Snyder and Palmer, 2004) we see that only few systems outperformed the heuristic consisting of choosing the most frequent sense as derived from SemCor (61.5%) and the results are better only by few points.

Obviously, the commonest sense is a notion that has to be considered not as an absolute indication of a specific sense but rather a statistical individuation of something highly dependent on the type and dimension of the reference corpus. In the current research, we can try to individuate the commonest sense by recurring to two sources of information:

a. the semantic layer of the Italian Syntactic-Semantic TreeBank
b. the internal order of the synsets in IWN and of the SemUs in SIMPLE-CLIPS.

The first source of corpus frequencies is only available for the experiment on the IWN database. The Italian Syntactic-Semantic TreeBank (ISST, Montemagni *et al*., 2003) consists of two sub-components: a generic and a domain-specific (financial) corpus, of about 215,000 and 90,000 tokens, respectively. The annotated material includes instances of newspaper articles, representing everyday journalistic Italian language. As far as annotation is concerned, the ISST has a three-level structure: two levels of syntactic annotation (a constituency-based and a functional-based annotation level) and a lexical-semantic level of annotation. In the ISST, sense annotation was performed manually using the ItalWordNet lexicon as a reference resource and the resulting annotation was used in the Senseval-2 and Senseval-3 lexical sample task (Calzolari *et al*., 2002 and Guazzini *et al*., 2004). Semantic annotation was performed by assigning a given sense number to each full word or sequence of words corresponding to a single unit of sense (such as compounds, idioms, etc.).

We carried out an experiment by using the list of frequencies extracted from the ISST; the list comprehends, for each row, the POS, the sense in the IWN lexicon, the frequency of sense in the corpus. The 249 occurrences of nouns in the CLEF2004 question collection was manually disambiguated (by one annotator), using as reference resource the same version of the IWN database already exploited in the annotation of the ISST. Then we compared the manually obtained results with the results we would have

[51] Descrizione del lexical sample task
[52] The two values respectively corresponding to the performance on the fine-grained and coarse-grained evaluation. The fine-grained evaluation is carried out by considering the original sense distinction in the reference resource (WordNet1.7). The coarse-grained evaluation was instead obtained by exploiting a list of grouped senses.
[53] Fine and coarse grained scoring.

obtained by exploiting the "most frequent sense heuristic" based on both the frequencies in the ISST and the first sense of the IWN and SIMPLE-CLIPS database.

Next tablee shows the percentage of overlap for the different PoSs: we obtain respectively 83%, 85% and 91% of correctly recognized nominal, verbal and adjectival senses in the CLEF2004 question collection. 9%, 48,1% and 33,3% of the senses were not available in the frequency list but the right sense in those cases was the first one for 87%, 64% and 100% of the times.

The results for nouns and adjectives are in line with the alternative method consisting in choosing the first sense in the IWN sense inventory: in this case, the senses were correctly disambiguated 86% of the times for the subset of nouns and 91% of the times for the adjectives. It is exactly the same result we get when considering the SIMPLE-CLIPS database (but with a higher number of not encoded SemUs, 12.4% of the total).

For verbs, however, the situation is not the same and the "first sense in IWN" method provides the worst results with only 64.4% of the correctly identified senses.

	Nouns	**Verbs**	**Adjectives**
# Occurrences	249	106	13
% occurrences most frequent in the ISST (according to IWN sense inventory)	83%	85%	91%
% occurrences corresponding to the first sense in SIMPLE-CLIPS	86%	56%	61.5%
# occurrences not covered by the SIMPLE-CLIPS SemUs	11%	17%	23%
% occurrences corresponding to the first sense in IWN	86%	64.2%	91.6%
# occurrences not covered by the IWN variants	2%	8%	0%
# occurrences not in the IST	9%	48.1%	33.3%
% occurrences not in the corpus corresponding to the first sense in IWN	87%	64%	100%

Table 12: results of the WSD on the CLEF-2004 test set

When disambiguating the senses in SIMPLE-CLIPS, it is worth remembering that, differently from IWN, the SemUs are not accompanied by a sense number. As a matter of fact, the only thing that allows the system to order the SemUs of the same lemma is their ID, which increases in a chronological way (the first encoded sense has a lower number in its ID). Also without the explicit instruction to do so, the lexicographer usually encodes first the most general/important/frequent sense, thus allowing the identification of a "first sense" in term of importance.

We have to highlight that neither lexicons have been built by taking into real consideration the linkage between lexical entries and corpus occurrences. The assignment of the role of first sense of the list is mainly based on the intuition of the lexicographer and on the order followed in the printed dictionaries used as source of information. Nevertheless, our study shows that, at least as far as nouns and adjectives are concerned, the first sense is almost always correspondent to what is needed in our specific task and testbde. Nevertheless, the results for adjectives would deserve further investigation by analysing a larger collection of questions, since the adjectival occurrences are too rare to be really representative.

The results of the two methods based on the corpus and on the order of senses are quite comparable. Nevertheless, when exploiting the IWN database, we decided to choose the disambiguation based on the commonest sense in the corpus since it seems capable of providing better results for verbs. For the lexical entries that were not annotated, we chose the first sense in the IWN lexicon. Since no corpus annotated according to SIMPLE-CLIPS is currently available, when considering the experiments on that lexicon we were obliged to rely solely on the order of the SemUs.

The result of the disambiguation will be used in all the modules of LR exploitation. Since we can expect that a more complex WSD system could provide better results, the output of the various modules should be conceived as a hypothetical lower bound of the performance of the system.

4.3.3.3 SEMANTIC SALIENCE OF THE KEYWORDS

In (4.3.3.4.6) we introduced a method to assign a relevance score to each keyword of the question, mainly based on the recognition of the PoS. Nevertheless, semantic salience and degree of term variation seem to be something that cannot be fully determined by taking into consideration only the part of speech of the keywords.

In order to understand which are the most important keywords in the question, we will evaluate the impact of the exploitation of information of a semantic nature; we said (in 4.3.3.3.) that the opposition between abstract and concrete entities does not seem to play any role in the selection of the keywords. Differently, we will show that the evaluation of the specific/generic opposition can be exploited even if only for a particular type of question.

4.3.3.3.1 General Vs specific Nouns

As we already introduced in 2.5.2.2, Paşca and Harabagiu (2001) individuated in specificity the semantic feature that would help the system to determine the salience of a question keyword. According to this idea, very specific keywords should not be dropped from the query. (Paşca and Harabagiu, 2001) also show that, when the specificity is taken into account in the selection of the keyword, the number of the TREC-8 correctly answered questions increases from 133 to 151 (that can be considered an extraordinary result). We want to test the validity of such an assumption in answering the CLEF2004 questions introduced by the pronoun Quale.

The system can determine specificity of the keyword by assessing two measures (generally inversely related):

- the number of hyponyms of the corresponding concept (as done in Paşca and Harabagiu, 2001), i.e. the so-called branching factor (Devitt and Vogel, 2004);
- the number of levels in the hyperonym chain above the concept.

In the enhanced prototype, we want to test a measure of specificity that takes into consideration both information types, by counting them off-line and by storing them in the database dedicated to IWN. The counting of the levels has been facilitated by the fact that the hierarchies of nouns and verbs have been indexed with the technique described in (Mihalcea, 2002). Contrarily to what was done by (Paşca and Harabagiu, 2001), in the count of the hyponyms we also consider multiword expressions (this means that we count *casa discografica* among the hyponyms of casa[54]). Moreover, in the setting up of the experiment, some decisions were taken, concerning the consideration of multiword expressions (MWEs) and word sense disambiguation (WSD).

In fact, some question keywords should be considered not in isolation but rather as parts of poly-lexical units. This is true, for example, for *bomba atomica* in *In quale anno è stata lanciata la bomba atomica su Hiroshima?*, for *campo di sterminio* in *Dove si trova il campo di sterminio di Auschwitz?*, for *salto con l'asta* in *Chi è il primatista mondiale di salto con l'asta?* etc. We counted 16 multiword expressions in the CLEF-2004 test set[55]. Most of these MWEs are listed among the lexical entries of the IWN database, while only one can be found in the SIMPLE-CLIPS lexicon, where globally only 13 MWEs (*fenomeno atmosferico*, *evento cognitivo*, *stumento musicale* etc.) were introduced as dummy entries to help categorization of homogeneous sets of senses. We decided to consider MWEs in the count of the hyponyms and not their individual parts since this strategy seemed more semantically founded. This means that in *Quando c'è stato un colpo di stato a Cipro?* we consider the nominal multiword expression *colpo di stato (*that has no hyponyms in IWN) as the keyword and not the two keywords *colpo* and *stato* with respectively

[54] Where the sense of casa is the fifth in IWN, corresponding to the WN1.5 synset {firm, house, business firm}.
[55] It is obviously not a fixed number since the distinction between what is a multiword and what is not in not sharp.

their 20 and 2 hyponyms (in the IWN db). In the same way, in the question *Qual è l'unità di misura della frequenza?* we consider *unità di misura* as the keyword (with its 129 and 72 hyponyms respectively in IWN and SIMPLE-CLIPS) and not *unità* and *misura* (respectively 6 and 41 hyponyms in IWN)[56]. 9 of the 16 poly-lexical keywords are not contained in LR: in those cases, if they were partially compositional (like for example *bomba atomica* and *salto con l'asta*) we decomposed them and counted the hyponyms of their parts, while if they were not compositional (like *acido salicilico*) we did not provide any number of hyponyms[57]. Obviously, in order to use this kind of information in a real system, the system itself should be provided with a module for the recognition of multiword expressions in the question text. This functionality could be incorporated in the parser or it may be thought to simultaneously exploit the parser output and the MWEs repository provided by IWN. Our system does not benefit from such a module but we think that its functioning can be simulated in order to obtain the input we need.

We decided to discard from the total number of hyponyms the SemUs that in SIMPLE-CLIPS are Proper Nouns. As we already explained, in IWN instances are treated differently from common nouns since they are connected to the class by means of the BELONG_TO_CLASS relation and not via the normal IS-A. This is not true for SIMPLE-CLIPS, where Proper Nouns, Nouns and Verbs are all gathered under the same nominal hyperonyms. It is useful for the system, when exploiting the resource in order to assess the level of specificity of the lexical item, to avoid counting the Proper Nouns among the other hyponyms since, in our opinion, they do not determine a major level of specificity. As a matter of fact, the number of instances is often motivated only by the choice to cover a specific area of the lexicon or not: the concept *petroliera* (oil tanker) is intuitively a quite specific concept even if, for particular applicative exigencies, the lexicographer may want to add a long list of names of oil tankers directly in the resource. Again, this decision is different from the one adopted in the experiment described in (Paşca and Harabagiu, 2001), where *city*, that in WN1.7 has only hyponyms of type instances, is indicated as a general term. When considering SIMPLE-CLIPS, we had to discriminate between hyponyms of the type "common noun" and of the type "proper name", since no distinction is made at semantic relation level.

As a threshold, an average measure of 10 hyponyms and 4/3 levels for both lexicons was established.

According to the sense inventory provided by IWN, there are 49 questions in the CLEF-2004 test set with keywords with more than 10 hyponyms. Diversely, by taking into consideration the hyperonymy links in Simple-CLIPS, we counted only 10 questions with keywords with a similar number of hyponyms.

Finally, we tried to extract the generic nouns by considering the co-occurrence of two conditions:

i) at least 10 hyponyms (all levels),
ii) a maximum number of 4 levels in the hyperonym chain

[56] In the CLEF-2004 test set we recognized only one verbal multiword expression: *essere in grado* (to be able).
[57] The level of "compositionality " of a mwe is more a continuous than a discrete measure. There seems to be a kind of continuum where the level of cohesion of single lexical items varies. For this reason, understanding what is a multiword and what is not is not easy and is one of the most challenging tasks of the discipline. In this analysis we simply decide to isolate the expressions that we thought would be useful to be treated as a unique lexical item.

A manual analysis of the CLEF2004 questions shows that usually the generic noun is the answer type term while the rest of the question is often well specified. Probably, this is due to fact that, generally speaking, the answer type term substitutes something that we cannot exactly specify (that is why it is the object of the question, the word we are trying to determine). Obviously, the "specificity rule" has to be considered as that works in the majority of cases but not always: in the case of CLEF2004question#176, for example, we see that the ATT "partito" should not have been sent to the Search Engine even if it cannot be considered an actual generic term. Given this situation, it is important to understand which ATTs are not worth sending to the IR module and we know that this is especially crucial for questions introduced by the pronoun *Quale*.

The next table comprehends all the questions in the CLEF2004 test set that are introduced by the interrogative pronoun *Quale*. Between brackets we have indicated (for the two computational lexicos): i) the number of hyponyms, ii) the depth of the taxonomy of the ATT. A Y/N field follows each question, corresponding to the necessity of submitting the ATT to the Search Engine in order to retrieve pertinent paragraphs. In this evaluation, we also applied the above said method of disambiguation (we also indicated when the assigned sense was not the "correct" one, it happened 5 times in the 16 questions when exploiting SIMPLE-CLIPS).

Question#4: Qual è *l'unità di misura* di frequenza	(IWN: 129, 3) (SIMPLE-CLIPS: 70, 0)	No
Question#6: Qual è il *nome* battesimo del giudice Borsellino	(IWN: 0, 7, "unità_linguistica" taxonomy) (sense not in SIMPLE-CLIPS; the "wrong sense": 5, 6, template MetaLanguage)	No
Question#11: Qual è la *città* sacra per gli Ebrei	(IWN: 7, 6) (SIMPLE-CLIPS: 4, 4)	Yes
Question#23: Qual era il *nome* di battesimo di Hitler	(IWN: 0, 7, "unità_linguistica" taxonomy) (sense not in SIMPLE-CLIPS; the "wrong sense": 5, 6, template MetaLanguage)	No
Question#26: Qual è la *capitale* della Russia	(IWN: 0, 7) (SIMPLE-CLIPS: 0, 5)	Yes
Question#28: Qual è il *titolo* del film di Stephen Frears con Glenn Close, John Malkovich e Michelle Pfeiffer	(IWN: 4, 6) (SIMPLE-CLIPS: wrong sense: 23, 4-Template Convention)	No
Question#31: Qual è la *professione* di James Bond	(IWN: 20, 4) (SIMPLE-CLIPS: 0, 6)	No
Question#50: Qual è il *quotidiano* italiano più letto	(IWN: 0, 8) (SIMPLE-CLIPS: 0, 3)	Yes
Question#52: Qual è un *ingrediente* base della cucina giapponese	(IWN: 0, 3) (SIMPLE-CLIPS: wrong sense: 0, 2)	No
Question#60: Qual è la *capitale* del Giappone	(IWN: 0, 7) (SIMPLE-CLIPS: 0, 5)	Yes
Question#91: Qual era lo scopo della prima azione sostenuta da Greenpeace?	(IWN: 6, 2) (SIMPLE-CLIPS: 0, 1)	No
Question#94: Qual è un *fattore* di rischio per le malattie cardiovascolari	(IWN: 0, 3) (SIMPLE-CLIPS: 0, 1)	Yes
Question#95: Quale è la *categoria* professionale più a rischio di cancro ai polmoni	(IWN: 21, 2) (SIMPLE-CLIPS: wrong sense 0, 5)	No
Question#145: Qual è la *sigla* dell'Esercito di liberazione del popolo sudanese	(IWN: 4, 6) (SIMPLE-CLIPS: 0, 6)	No

Question#176: Qual è il *partito* di Charles Millon	(IWN: 5, 4) (SIMPLE-CLIPS: 0, 4)	No
Question#196: Qual è la *valuta* irachena	(IWN: 37, 4) (SIMPLE-CLIPS: 12, 3)	No

Table 13: questions in the CLEF2004 test set introduced by the interrogative pronoun *Quale*

If we analyse the questions, we see that two tendencies seem to emerge:

i) meta-linguistic ATTs should never be present in the query (*nome*, *titolo*, *sigla*, *abbreviazione* etc.).

ii) generic, vague terms often do not appear in the answer. Intuitively, terms like *ingrediente*, *professione*, *unità di misura* etc. can be considered generic terms, because we expect them to categorize a certain number of things and should also be quite high in the hierarchies.

As regards the first exigency, we can exploit LRs to recognize the "meta-linguistic" ATTs, that are categorized: i) under the node {unità_linguistica} in IWN (the TC LANGUAGEREPRESENTATION is too generic and includes terms that cannot be considered metalinguistic, for example *quotidiano* of CLEF2004question#50) and ii) under the Template METALANGUAGE in SIMPLE-CLIPS.

As regards the specificity option, we see that in this case IWN (even with two exceptions) is able to provide a useful support in recognizing the lexical items that should or not be sent to the Search Engine. By exploiting IWN, all the cases of ATT expressing metalinguistic information were correctly recognized. SIMPLE-CLIPS failed 4 times in recognizing general terms and once in recognizing metalinguistic word meanings. In chapter 5 we will analyse the reasons behind these failures.

As a result, in the enhanced prototype we decided to implement a module that does not send generic ATTs to the Search Engine. This strategy determines that in the case of CLEF2004-question#4 *Qual è l'unità di misura di frequenza?* and of CLEF2004-question#196, *Qual è la valuta irachena,* only the noun *frequenza* and the adjective *irachena* are respectively sent to the IR module (we will later see that LR also allow the system to submit dynamic queries made with the hyponyms of the ATT, restricting an exaggerated recall in this way).

4.3.3.4 ENRICHING THE ANSWER TYPE TAXONOMY WITH LEXICO-SEMANTIC INFORMATION

In paragraph 4.3.2.2, we introduce the Answer Type Taxonomy, i.e. the hierarchy of expected answer types. We saw that the hierarchy exploited within the "baseline prototype" contains only 22 nodes, while the analysis of the questions of the tenth TREC campaign induced the identification of more than 40 nodes. The analysis of the results of the baseline prototype confirms that such a coarse classification is not enough to handle the numerous types of expected answer (4.3.2.8).

As we have already learned from (Pasca, 2003; Voorhers, 1999) and as the results of the baseline prototype show, the most problematic cases are represented by questions introduced by the interrogative

adjectives and pronouns *Che* and *Quale*. In the capacity of interrogative adjective, *Che* is ambiguous when interpretating the selection of individuals and classes: when it is used to ask about an individual to be chosen from a group it overlaps, especially in North Italy, the interrogative element *Quale* (Renzi, 1995). For both, the same consideration is valid: generally, the AT refers to the semantic type of the noun modified by the interrogative adjective (the answer type term). When we presented the Answer Type Taxonomy exploited in the baseline prototype, we showed that some very frequent lexico-syntactic patterns introduced by *Quale* were inserted in the clusters of some ATs. This is true, for example, for some very frequent types of Location, such as *città*, *paese* etc.. In this case, there is not an actual need to exploit information stored in lexical-semantic resources: when we find questions like CLEF2004-question#29 *Quale paese confina a nord con il Canada?* (What country is bounded on the north by Canada?) and CLEF2004-question#184 *In quale città la Mosella incontra il Reno?* (In what town does the Mosel meet the Rhine?), the simple pattern matching on the pattern *Quale* + *paese* and *Quale* + *città* is enough to guarantee the correct derivation of the type of expected answer. Even if in this case the baseline prototype was sufficient to derive the expected answer type, it is clear that a more general strategy to handle this type of questions is required. As a matter of fact, the ATT can be anything: remaining in the "location" type of answer, a question can ask about a city but also about a village, a specific address, a neighbourhood, an expanse of sea, a border between two countries etc. A good example of the variety of the situation is the translation of some of the TREC-10 questions asking about location:

Qual è l'indirizzo della Casa Bianca?
In quale oceano sono le Isole Canarie?
Qual è il lago più profondo degli Stati Uniti?
Quale monastero fu saccheggiato dai vikinghi nel tardo ottavo secolo?
Quale pianeta ha il più forte campo magnetico?
Quale è la stella più brillante?
In quale contea della California si trova Modesto?
Qual è la capitale dello Zimbabwe?
Quale stretto separa il Nord America dall'Asia?
Di che penisola fa parte la Spagna?
In quale emisfero sono le Filippine?
In quale provincia francese viene prodotto il cognac?
Nel tardo 700 quale colonia era popolata da prigionieri inglesi?
Qual è la più grande faglia vicino al Kentucky?
Quale parco nazionale si trova nello Utah?
Quale porto sovietico è sul Mar Nero?

Indirizzo (address), *oceano* (ocean), *lago* (lake), *monastero* (monastery), *pianeta* (planet), *contea* (county), *stretto* (strict): these questions are not introduced by the interrogative adverb *Dove* (*Where*), but they are indeed used to ask about something that can be classified as a location (according to an ontology of types)[58]. The AT can thus be Location and the process of answer identification can be even more sure and simple if we have at our disposal a Named Entity Recognizer capable of detecting more fine-grained classes of entity.

[58] The questions introduced by *Dove* and the ones of the type *Quale* + *location* are not the same, there are important differences concerning the specificity of the expected answer. Nevertheless, it is important to for the system to trigger the same type of methodology in the processing of the answer.

It is here that language resources would be useful, in helping the system to address all these different word meanings towards a common type of expected answer. In many cases, in fact, the semantic type of the noun modified by the interrogative adjective is the only thing that tells us that we have to look for a named entity of a given type in the candidate answer. For this reason, the synsets that lead the taxonomies concerning location were linked to the corresponding AT in the ATTaxonomy.

As we already pointed out in 4.3.2.8.1, the questions introduced by imperatives such as *nomina* (name), *dimmi* (tell me), *dammi* (give me) and by the patterns "*Qual è il nome*…" and "*Come si chiama*…" cannot be answered without recurring to sources of information that help the system to analyse the semantics of the ATT. Examples of this type of question are *nomina una compagnia petrolifera, Come si chiama la compagnia di bandiera tedesca?, Come si chiama la casa discografica di Michael Jackson?* Etc.

It is worth noticing that in the case of questions introduced by the interrogative forms "*come si chiama….?*" and "*qual è il nome di…?*", the derivation of the AT is not always an easy task. Usually, in fact, these forms are used to obtain the name and surname of a person: in the case of CLEF2004question#121: *Come si chiama la moglie di Kurt Cobain?*, the system analyses the ATT *moglie* and derives the correct AT HUMAN and the corresponding Named Entity Type, i.e. Person. Also the ATT *pilota* of CLEF2004question#14 is categorized as a Human in both the LRs but the expected answer is not a human name but rather a "specific type" of *pilota*, what we could call a hyponym. The adopted strategy consists in testing the number of the ATT classified as human and, in the case of a plural, triggering a specific answer detection strategy (the dynamic query we describe in 4.3.3.4.6). The AT remains the same but the strategies adopted to identify the answer are different.

There are other types of question whose expected answer types are even less explicit. Consider the Italian correspondents of the TREC-10question#899: What is the life expectancy for crickets? (*Quanto vive in media un grillo?* and *Qual è l'aspettativa di vita di un grillo?*). The Answer Type is in both cases a temporal expression but how could the system understand this? In the first case, *Quanto vive in media un grillo?*, LRs should provide the system with the notions that allow it to discriminate among the various interpretations deriving from the same syntactic form *Quanto + verb + subject?* The system should be able to capture the temporal shade behind a specific sense of *vivere*, that is completely different from the meaning humans are immediately able to interpret when they are asked to answer the question *Quanto mangia in media un grillo?* (i.e. a quantity). For the second case, *Qual è l'aspettativa di vita di un grillo?*, the system should be provided with the possibility of capturing the semantic deriving from the modification of the noun *aspettativa* (expectancy) with the noun *vita* (life). Also verbs like *vivere*, *costare*, *ammontare*, *durare* etc. should be interpreted by recurring to the Answer Type Taxonomy, thus allowing their (partial) understanding by the system. The same can be said for the questions where the ATT is an adjective, like *Quanto è alto?*, *Quanto pesa?* etc. It would be important to have a method of deriving their Answer Type in a systematic way without recurring to ad hoc rules like the ones we encoded in the baseline prototype. We think that the most promising strategy consists in exploiting the Top Concepts and the Semantic Types of the Top Ontologies of the two lexicons.

4.3.3.4.1 Final architecture of the AT Taxonomy

As we already explained in 4.3.2.2, two disjointed types of expected answer were identified: the first type consists of the answers referring to single factual information (a person's name, a specific location, a length expressed in meters etc.); the second type refers to more complex answers, describing a series of events, explanation, reasons etc. The highest nodes, FACT and DESCR refer respectively to these two most general categories. We also showed that 22 types of expected answer could be determined by recurring to stem analysis and pattern matching. So, we decided to recur to the strategy adopted in the FALCON system (Harabagiu *et al*., 2001; Paşca, 2003) in order to make ItalWordNet and SIMPLE-CLIPS sustain the exigencies in terms of node representation. The nodes in the ATTaxonomy have been projected on the branches of the ItalWordNet taxonomies[59] and on the SemUs of SIMPLE-CLIPS. As we said, we recognized a possible set of expected answer type composed by about 40 Answer Types. Nevertheless, the analysis of the testbde shows that, even if some major categories can be recognized and defined, the set of possible expected answers is virtually infinite, it has no clear boundaries and depends on the level of specificity and of informative power one would expect from an automatic system. The issue is, having the possibility of relying on information in semantic computational lexicons, what is the "right" set of nodes that should be inserted in the Answer Type Taxonomy? What is the representative modality that would allow the system to handle and answer the majority of questions? We decided to host in the taxonomy the ATs referring to the following cases:

1. semantic types corresponding to Named Entity categories
2. semantic types that can be individuated by recurring to specific strategies
3. very frequent types of expected answer

The first type of ATs includes for example the nodes CITY, LENGTH, WEIGHT, SPEED, etc. The strength of this type of ATs is that the answer is something that can be recognized in the text as belonging to some Named Entity classes. The number of this type of ATs is obviously determined by the types of classes that can be actually recognized by the system. We preferred to consider all the NE classes that are plausibly recognizable by a good NERecognizer, so we added to the ATTaxonomy the node HUMAN, CITY, RIVER, MOUNTAIN, COMPANY etc. The hierarchical organization of the ATs allows us to freely decide to exploit a more or less underspecified named entity class without having to restructure the taxonomy.
The second type of ATs refers to a AT whose identification can trigger specific rules that allow the system to find the candidate answers. For example, the detection of the ATTs ragione, causa, spiegazione, motivo etc. can trigger the heuristic that helps the system to find an explanation to something, the same that should be activated in the case of Perché (why) questions (that have AT REASON).

[59] The ItalWordNet tool developed at ILC-CNR was used to encode both the ATTaxonomy and the links to IWN.

The third type of AT refers to the most difficult types of question. As a matter of fact, the analysis of the test-bed shows that there is a number of questions whose expected answers are not named entities nor something that can be gathered under the node DESCR of the ATTaxonomy (i.e. an EXPLANATION, a REASON or a DEFINITION). They are the ATTs described as "too specific" in (Paşca, 2003, pag. 68)60. This is true, for example, for questions like CLEF2004question#31 Qual è la professione di James Bond?[61] and CLEF2004question#48 A quale pena è stato condannato Pietro Pacciani per i delitti del Mostro di Firenze?[62]: in the cases like these, the answer is a nominal concept that should be listed as lexical entry in the reference resources and that can be sought among the hyponyms of the ATT. A possible strategy for solving these cases thus consists in exploiting the hyperonym chain not bottom-up (to understand the type of expected answer) but rather top-down (to use the set of hyponyms of the ATT). In this sense, the actual presence of a specific node of the type profession or penalty is obviously not necessary, since the question has to be analysed not by abstracting from the ATT but rather by exploiting its subsumed concepts. Nevertheless, the analysis of the TREC and CLEF question collections highlights some recurring types of expected answer that can be classified accordingly to more general concepts. For this reason, many ATs were added to the ATTaxonomy concerning entities like animals, garments, instruments, monetary units, units of measurement etc. The first utility in adding these nodes is that sometimes a more underspecified classification (like the one we obtain by individuating common ATs that gather different taxonomies) allows the system the more flexible search in wider classes. Consider, for example, the Italian correspondents of TREC10question#1011 What mineral helps prevent osteoporosis?, Quale minerale aiuta a prevenire l'osteoporosi?. The AT of this question is SUBSTANCE. If the system searched for the answer only among the hyponyms of the synset corresponding to the ATT minerale, it would be doomed to failure because in IWN the answer, {calcio 2}, is classified as a metal and not as a mineral. The same happens also in SIMPLE-CLIPS: calcio is classified as elemento (that has no hyperonym), mineral as sostanza. The two SemUs share only the Template (Natural_Substance) and the AT we imposed in the ATTaxonomy (SUBSTANCE). The situation can be resolved by exploiting the entire taxonomy subsumed by the AT SUBSTANCE, including also in this way the "right" synset and SemU. In the enhanced prototype, thus, we foresaw a two-step search firstly exploiting only the taxonomy subsumed by the ATT and then, if no match is returned, the whole taxonomy subsumed by the AT. Another useful function of these specific types of AT is the possibility of automatically keeping track of the ATs that most frequently occur in the QA practice, in order to be able to develop successful strategies to solve their cases. For this reason, we decided to add, at the output of the Question Analysis phase, a final statistics that records the number and types of the recognized ATs. It is also worth noticing that a similar automatic tagging of the question collection according to a given taxonomy of types may in future be useful in view of statistical, machine learning extension of the application. The final version

[60] We do not think it is a matter of specificity but rather of being or not an instance instead of a common nominal concept. Nevertheless, the distinction between instances and common nominal concept is not easy as well.
[61] What is James Bond's job?
[62] What penalty was Pietro Pacciani sentenced for the Florence monster murders?

of the ATTaxonomy comprises 43 nodes but it can be revised and improved with every new collection of questions that will provide new cases for the system to handle and analyse.
The final architecture of the ATTaxonomy is articulated in two layers:

- the first layer is constituted by clusters of lexical-syntactic patterns typical of specific types of question, already exploited in the baseline prototype. They are conceived to map different syntactic realizations into a same semantic representation.
- The second layer is represented by the semantic articulation of the patterns: some ATTs are linked to the synsets and the SemUs of the two lexicons, and so become the roots of the taxonomies that roots of the taxonomies that collect senses revealing specific Answer Types.

It is important to remember that the final configuration of the taxonomy has been designed by working exclusively from scratch and by organizing bottom-up types of expected answers as they resulted from a manual analysis of the question collections. Nevertheless, the final taxonomy has also been compared with a public available taxonomy, the one prepared by L. Ferro of the MITRE corporation (Ferro, 1999) on the TREC question collections[63]. The two taxonomies are quite similar, even if ours is a little bit more detailed (43 Vs. 33 nodes) in particular because it recognizes a higher number of ATs corresponding to NE classes.

4.3.3.4.2 Exploitation of the IWN hierarchies

When we try to project the Answer Types on the ItalWordNet taxonomies, we can see that often the ATs have to be addressed on scattered portions of the semantic net. For example, the node LOCATION of the Answer Type Taxonomy can be mapped on the synset {luogo 1 – parte dello spazio occupata o occupabile materialmente o idealmente[64]}, that has 52 first level hyponyms and that we can further organize with other (at least) 10 sub-nodes, such as:

- COUNTRY (mappable on {paese 2, nazione 2, stato 4}),
- RIVER, {fiume 1},
- REGION, {zona 1, terra 7, regione 1, territorio 1}, {superficie 1, area_geografica 1, area 1}
- etc..

These most specific nodes would be useful if the system had the possibility of recognizing correspondent Named Entity classes. We think that it is however a better strategy to maintain the taxonomy open to further improvements of the system. When, like in our case, no NERec is able to individuate such specific classes, the system can however exploit the hierarchical structure of the taxonomy to derive the more underspecified node (in this case, LOCATION).

[63] Available on line at URL www.trec.nist.gov
[64] *place 1- part of the space that can be ideally or physically took up.*

Identifying most specific nodes is also useful for another reason: they provide a major articulation of the taxonomies and can be conceived as anchors where scattered portions of the IWN semantic net can be attached. As a matter of fact, while some of the taxonomies that have these synsets as roots are led by the same superordinate {luogo 1}, there are others that are differently classified. In the case of {luogo 1}, we have a superordinate that circumscribes a large taxonomical portion that can be exploited in the AT identification. Nevertheless, the analysis of the TREC set of questions showed that many questions expect as an answer other types of location that are not classified as places in the resource (sometimes not even at ontological level); for this reason, we added four other sub-hierarchies to this area:

- CELESTIAL_BODY {mondo 3, globo 2, corpo_celeste 1, astro 1},
- BODY_OF_WATER {acqua 2 – raccolta di acqua}, {corso 2, corso d'acqua 1}
- BUILDING {edificazione 2, fabbricato 1, edificio 1 – costruzione architettonica}.

It is possible to see that sometimes more than a single synset has been mapped onto the node of the ATTaxonomy. For example, the AT BODY_OF_WATER is used to gather many ATTs, such as *mare* (see), *lago* (lake), *stagno* (pond) etc. In order to collect all these similar items, two different synsets ({acqua 2 – raccolta di acqua} and {corso 2, corso d'acqua 1}) have been linked to the same AT.

Fig. 47 gives an idea of the way the nodes of the ATTaxonomy are projected on the nodes of the IWN hierarchies: the circumscribed taxonomical portion includes the nodes directly mapped on the ATs, all their hyponyms (of all levels) and all the synsets linked to the hierarchy by means of the BELONGS_TO_CLASS/HAS_INSTANCE relation. As a matter of fact, while in the American WordNet the synsets of type instance are linked to their superordinates by means of the normal HAS_HYPERONYM relation (not distinguishing, in this way, classes from instance), in ItalWordNet the HAS_INSTANCE/BELONGS_TO_CLASS relation is used in these cases.

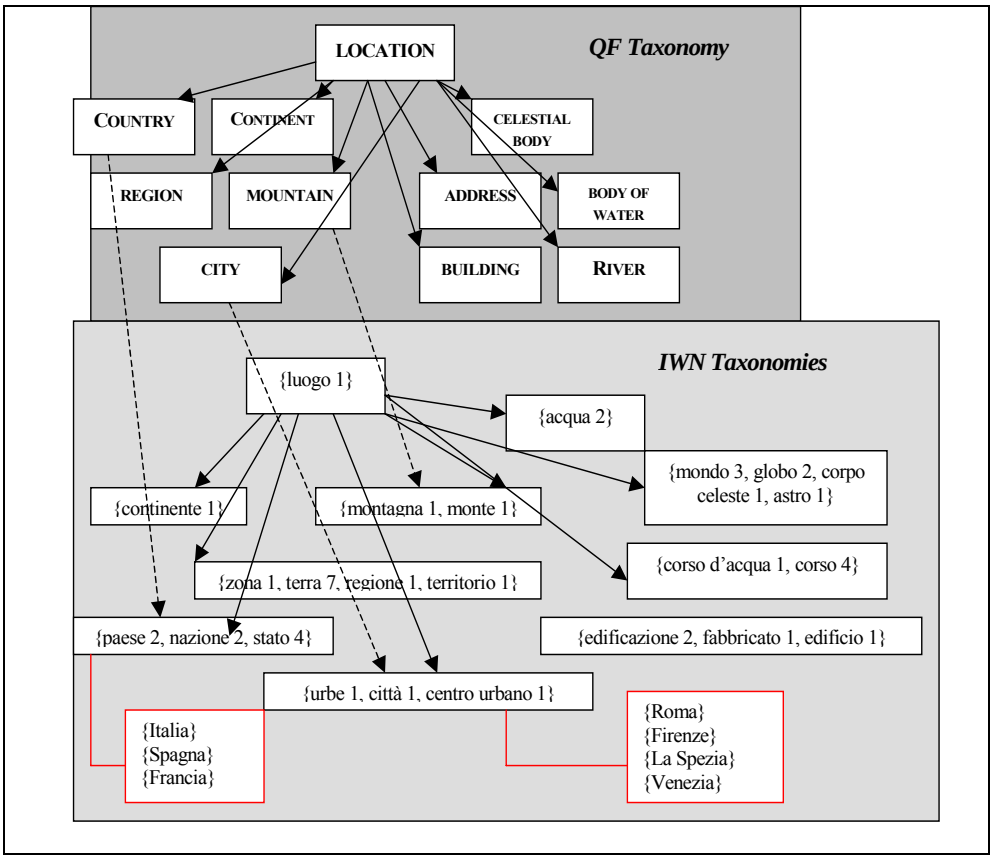


Fig. 47: Mapping the node Location of the ATTaxonomy on the lexical nodes of IWN

The ATTaxonomy and its links to the two LRs are the basis for a specific module of the system that retrieves the Answer Type of many questions of the type *Quale* and *Che*. In this way, we obtain the Answer Type CITY of question#3 (*In quale citta' si trova il carcere di San Vittore?*[65]), the AT ANIMAL of question#? (*Quale animale tuba*?) etc. This derived information will allow the system to filter out non-pertinent paragraphs (i.e. paragraphs not having any lexical entity respectively of type city and animal) during the answer extraction phase.

As we already explained, however, the strategies to single out the answer will be different: when a Named Entity class is linked to the node of the ATTaxonomy, the answer will be sought only among the named entity of the corresponding type present in the paragraphs (in the case of question#3, the Named Entity CITY). We can see that sometimes the answer is a non-lexical Named Entity, like in the case of the ATTs *velocità (Qual è la velicità raggiunta in volo da un Boing 747?*), *temperatura* (*Quale temperatura c'è al centro della Terra?*) and *percentuale* (*Quale percentuale d'acqua c'è nel corpo umano?*)[66]. All these expected answers (AGE, SPEED, DATE, WEIGHT, PERCENT, TEMP, LENGTH, COST) have been represented as sub-nodes of the QUANT AT in the Answer Type Taxonomy.

[65] *In what city is the San Vittore prison?*
[66] These questions belongs to the translation of the TREC-10 test set since no questions of this type were present in the CLEF-2004 question collection.

Differently, when no Named Entity class is linked to the ATTaxonomy node, like in the case of CLEF2004-question#? *Quale animale tuba*?, the searching routine will be restricted to those paragraphs containing an entity of the type indicated by the ATT (by searching among the hyponyms (of all levels) of the noun).

Some ATs gather not only more than one synset but also synsets belonging to different PoSs. An example is COST, which collects verbal and nominal synsets as is illustrated in Fig. 48. The same figure also shows the XML structure used to store the various ATs and the links to the resources.

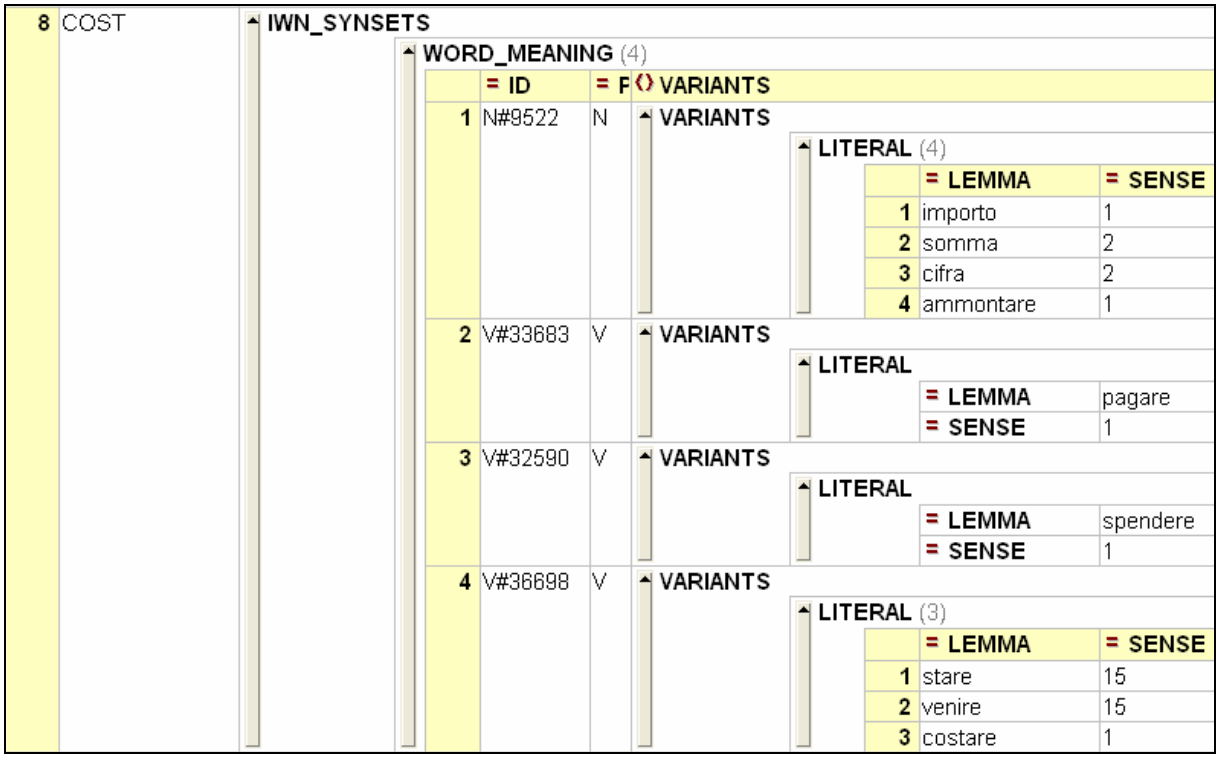


Fig. 48: sysets linked to the AT COST

The AT COST was already available in the version of the Taxonomy used in the baseline prototype, derived by exploiting an *ad hoc* rule on the patter *quanto+costare* (quanto costa X?). This new version allows the system also to recognize the expected answer in questions like *Qual è il costo di X?*, *Qual è il prezzo di X?*[67] but also *Quanto si spende per acquistare X?* and *Quanto si paga per X?*.

The result of the mapping procedure consists of 48 synsets that are now linked to the ATTaxonomy. This mapping covers about of the IWN taxonomies.

4.3.3.4.3 Exploitation of the Semantic Units in SIMPLE-CLIPS

[67] For what regards the ATT *prezzo* (cost), we prefered to link the QF to the higher synset {importo, cifra, somma and ammontare} from which it can be easily derived.

In the ATTaxonomy file a specific element is dedicated to the link between each AT and the corresponding SemUs in SIMPLE-CLIPS. The information is not an integration of the linking mechanism to IWN but rather an alternative that shows how the same methodology of AT derivation can work also by exploiting a different language resource. 73 SemUs are now linked to the ATTaxonomy. The number of SemUs directly mapped to the ATs is greater than the IWN synsets because what in IWN is gathered in the same synset, in SIMPLE-CLIPS is distributed in different SemUs. In the next paragraph, we will see that some ATs have been linked not to a specific SumU but to an Ontological Template.

4.3.3.4.4 What role for top ontologies?

A different way to group the lexical items of LRs together would be to recur to the IWN and SIMPLE-CLIPS Top Ontologies. The idea is interesting (Bertagna, 2003) because Ontologies classify the lexical content of LRs in wider portions, thus potentially allowing a more coarse-grained overlap on the nodes of the ATTaxonomy. We will see, however, that Top Ontologies do not seem to easily support the exigencies of AT derivation since mismatches between Top Concepts and Answer Types have to be resolved at a fine-grained level, i.e. the lexical one.

For what regards IWN, the way in which the ontological information is projected on the lexical nodes would allow us to select and circumscribe wide lexicon portions, kept together by:

i) the links between the monolingual database and the ILI portion hosting the Base Concepts,
ii) the links between the Base concepts and the TO,
iii) the ISA relations linking the synset corresponding to the Base Concept to its conceptual subordinates of *n* level, down to its leaf nodes.

The EWN Top Ontology, however, doesn't seem really suitable for determining AT, since we need more fine-grained distinctions to better adhere to the requirements of the task. In the case of questions about Location, for example, we can extract all the synsets belonging to the Top concept Place. But only the ATs Country, Region, Mountain, Continent, City and Body_of_water can be projected on this wide category, while River, Celestial Body and Building belong to other ontological portions (River and Celestial_Body are classified as Object/Natural while Building as Artefact/Building/ Object). The problem is that the Top Concepts Object and Artefact are too generic and not discriminating in the selection of the lexical items that can be used with the function of "places". For this reason they cannot be used to select the lexical area pertinent to the respective ATs, that could be selected by recurring to more discriminating lexical nodes such as {fiume 1}, {mondo 3, globo 2, corpo_celeste 1, astro 1},{edificazione 2, fabbricato 1, edificio 1}.

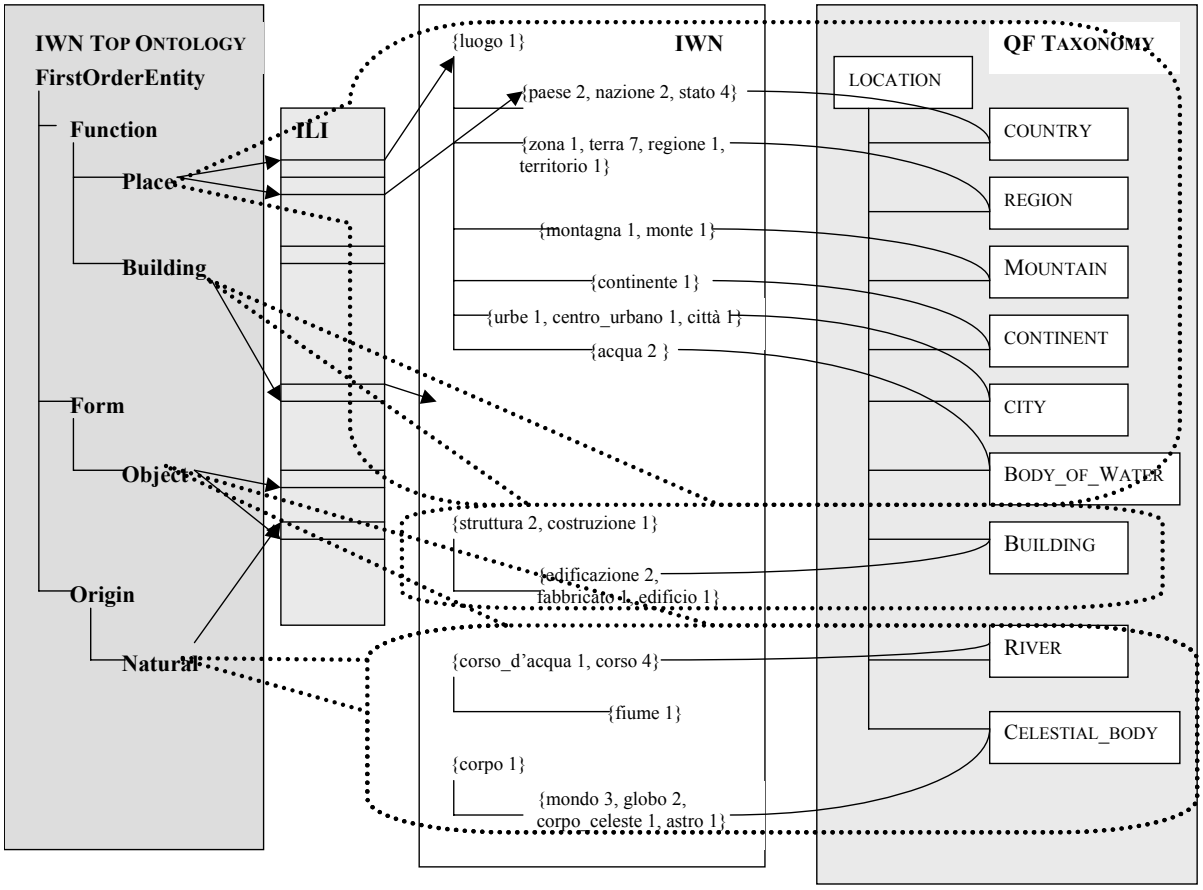


Fig. 49: Projection of the nodes about the AT Location on theTCs of the EWN TO

Other example of this type of mismatch between the aim of the task and the way the content of LR is organized is evident when we want to semantically interpret verbal and adjectival ATTs. The practice of linking the synsets of the semantic net with the node of the ATTaxonomy shows that it is very uncommon to find a semantic representation that complies with the required interpretation. Differently from the case of the synset {costare, stare, venire} (that is classified under the node Possession of the Top Ontology), nothing in the ontological classification of the synset {durare} (classified under the Top Concept Static) provides any clues about its fundamental dimension of meaning (the temporal one).

Thus, the exploitation of the Top Ontology nodes cannot be the default methodology for the selection of the relevant synsets.

On the contrary, establishing links between the ATTaxonomy and the ontological structures of the lexicon would seem to be the recommended strategy in the case of use of SIMPLE-CLIPS as reference resource. As a matter of fact, the SIMPLE-CLIPS ontology is more detailed than the IWN one (157 Templates Vs 68 Top Concepts in IWN) and can be exploited to select and circumscribe rather homogenous subsets. Moreover, differently from what happens when exploiting the synsets in IWN, there are some ATs that cannot be efficiently mapped onto any SemU. This is true, for example, for the AT BodyPart: while there is a correspondent synset in IWN ({parte_anatomica, parte_del_corpo}), in SIMPLE-CLIPS there is no SemU that organizes the body parts. What can be exploited is however the Semantic Type Body Part. The same happens for the AT HUMAN GROUP: in SIMPLE-CLIPS there is no sense of gruppo that specifically

covers the human groups, so we have to exploit the corresponding Template (HUMAN GROUP). Paradoxically, the ontological classification is in these cases more specific than the organization supported by the lexical items. In these two cases, specific rules are added to map the AT to the Templates: when the ATT is "parte_del_corpo", the system detects the BodyPart AT and, at the same time, searches among the SemUs classified as BodyParts to find the candidate answer.

Nevertheless, it is not possible to only exploit the Ontology instead of the lexical entries. In the majority of cases, in fact, the information in the Ontology is too general and abstract. To give an example, Fig. 50 shows a detail of the Simple Ontology dedicated to the Location node. The nodes of the ATTaxonomy overlap with the nodes of the SIMPLE Ontology (even if the relation between the semantic types and ATs is not biunique) but we encounter the same problem which emerged with the IWN TO, i.e. too generic Templates that do not allow us to completely rely on the ontological information to classify semantic content with respect to AT representation. As a matter of fact, Celestial_Body doesn't overlap with the Templates concerning Location, because the planets, the stars and in general the bodies of the sky are classified as Concrete_Entity, which is too generic to be useful. In order to map Celestial_Body to SIMPLE, it would be necessary to manually select a common and shared hyperonym in the lexicon (in this case, the SemU corpo).

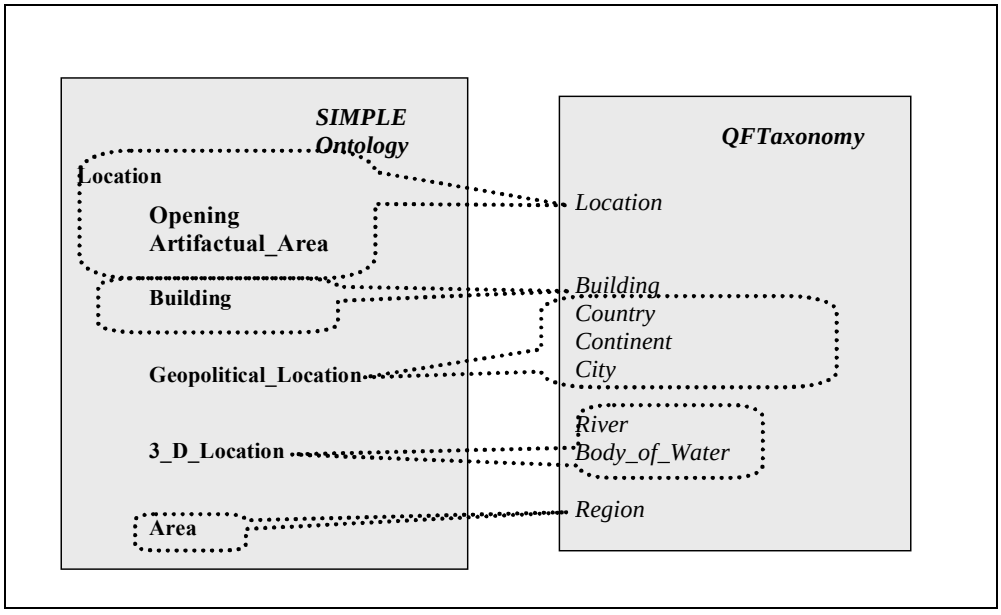


Fig. 50: the branches of the SIMPLE Ontology dedicated to Location and the ATTaxonomy

The experiment carried out by linking the two ontologies on the ATTaxonomy demonstrates that, when possible, it is better to use more fine-grained information at lexical level. Nevertheless, the Ontology of SIMPLE-CLIPS provides a useful alternative classification when an appropriate SemU cannot be mapped to the nodes of the ATTaxonomy.

4.3.3.4.5 Type *Taxonomy*

No link has been established between the ATTaxonomy and the taxonomical portion with root *tipo*, *sorta* etc. because it would be very vague. If we look at two questions of the TREC competition we can see that in the case of questions of the type "quale tipo/sorta/genere/ .." what really disambiguates the question is the modifier of the "type" word:

TREC1999question#1368: *What type of polymer is used for bulletproof vests?* (Quale tipo di polimero è usato per i giubbotti antiproiettile?)

TREC1999question#1376: *What kind of gas is in a fluorescent bulb?* (Che tipo di gas è contenuto in una lampadina fluorescente?)

For this reason, in the case of this type of question, a rule was foreseen in order to consider the complement of the "type" word as ATT. All the variants of the IWN synset {tipo, sorta, fatta, genere, specie, forma, qualità} were used to constrain the rule. When the modifier of the "tipo" word is an adjective, two strategies are foreseen:

i) if the "tipo" word and its modifier are already present as such in the language resource, the matching entry is used to exploit the available hyponyms. This is the case of CLEF2004question#42: *In quale genere musicale si distingue Michael Jackson?*

ii) if no multiword of the required type is encoded in the LR, then the adjective is analysed and the ATT becomes the noun linked to it by means of a SRDenominalAdjective relation (for SIMPLE-CLIPS) or of a PERTAINS_TO relation (in IWN).

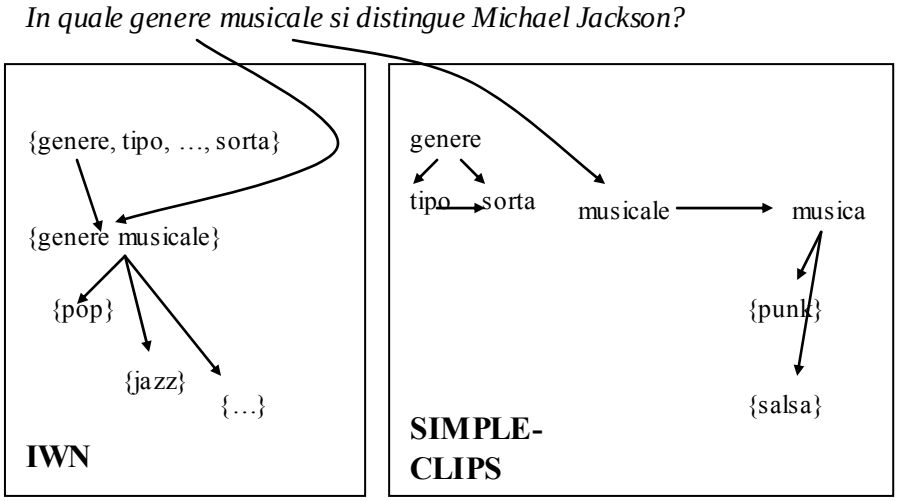


Fig. 51: strategies to handle questions of the type "quale sorta/genere/../tipo…?"

This strategy and its exception had to be adopted because, under the ontological point of view, the way in which the "type" taxonomy in IWN and SIMPLE-CLIPS is organized is very problematic and inconsistent. We will discuss this problem in 0.

4.3.3.4.6 Semantic feedback on the query formulation module: dynamically created queries

In 2.5.2.1.3 we introduced a strategy adopted in the FALCON system (Paşca, 2003) consisting of creating dynamic queries in the case of questions of the type "What type/sort/specie of..?". In these particular cases, the query is formulated by iteratively adding to the keywords of the query all the hyponyms of the ATT (in place of the ATT itself). The "philosophy" behind this strategy is very interesting from our point of view because it is based on the assumption that LRs can be used not only to understand and lexically expand the question but also as a repository of possible answers. Such an assumption also presupposes that the semantics of lexical items in the computational lexicons is enough to operate a total or partial match with the semantics of the sought entity.

An important difference between the methodology described in (Paşca, 2003) and the one we implemented in the enhanced prototype is the type and number of cases on which we decide to apply the dynamic query strategy. As a matter of fact, in the FALCON system, the dynamic query is created and submitted to the IR module only in case of questions with the form "What type/sort/specie of..?" (corresponding to the Italian "Quale tipo/sorta/specie di…?"). In the enhanced prototype, we decided to adopt the same strategy also in case of questions of type Quale+noun. The decision is due to the observation that it is not always easy to individuate a sharp difference in the use of the two superficial types of question: the translation of the example presented in (Paşca, 2003), What type of flower did Van Gogh paint? (Che tipo di fiore dipingeva Van Gogh?) can be reformulated in What flowers did Van Gogh paint? (Che fiori dipingeva Van Gogh?). In both cases, what the system should do is enriching the query with the hyponyms of the ATT flower, in order to generate the answer sunflower when it is submitted with the other keywords. The same strategy seems to be adapt in case of CLEF2004question#127, Quale animale tuba?, where the answer could be found by submitting *colombo* (pigeon) together with the keyword *tubare* (to coo). In that case, however, it is less sure that the alternative form *Quale tipo di animale tuba?* would be well formed and acceptable.

It is important to remember that *an X is a kind/type of Y* is indicated in (Cruse, 1986) as the diagnostic test for *taxonymy*, a sub-specie of hyponymy (*a spaniel is a kind of dog* Vs *?A waiter is a kind of man*). Cruse remembers that it is difficult to discover invariable semantic properties that differentiate taxonyms from all other hyponyms. It is possible, however, to recognize some clues: for example, (Cruse, 1986) says that hyponymy seems more suited for *nominal kind* while taxonymy more adapted for *natural kind* terms. Nominal kind terms (Pulman, 1983) are lexical entries that can be defined by encapsulating a

syntagmatic modification of their superordinates in the typical pattern genus+differentia (*stallion:male horse*). This type of terms are generally connected to their superordinates by means of a generic hyponymy relation. The nature of the greater specificity of natural kind terms relative to their superordinate remains on the contrary obscure: *horse* is a kind of *animal* but there is no modification of *animal* that can yield an expression equivalent to *horse* in the way that male *horse* is equivalent to *stallion*. An account of what sort of animal a horse is would required an encyclopaedic definition of indeterminate size and complexity.

What is the reason behind the decision to exploit, in the FALCON system, hyponymy relation only in case of questions of type "What type/sort/specie of..?"? We do not think that the reason can be the conviction that the expected answer is usually what can be referred to as a taxonymy of the ATT. As a matter of fact, in the FALCON QA system, thePrinceton WordNet is used and in that lexicon no specific encoding of taxonymy is envisaged (so only hyponyms can be retrieve by the system). It is thus not clear why questions with form "Quale tipo/sorta/specie/genere.." should be treated with dynamic query generation differently from all the other questions asking about some specification of the Answer Type Term.We thus decided to apply the same methodology also to the questions of the testbed with the form *Quale+ATT*, *Come si chiama+ATT*, *Dimmi il nome di+ATT*, *nomina un ATT* etc.:

q_4 Qual è l'unità di misura di frequenza?
q_9 Quale incarico ricopre Ariel Sharon?
q_14 Come vengono chiamati i piloti suicidi giapponesi?
q_18 Che lingua si parla in Germania?
q_24 Che moneta si usa in Germania? marco + Germania
q_31 Qual è la professione di James Bond?
q_42 In quale genere musicale si distingue Michael Jackson? Pop + Michal Jackson
q_48 A quale pena è stato condannato Pietro Pacciani per i delitti del Mostro di Firenze?
q_52 Qual è un ingrediente base della cucina giapponese?
Di quale nazionalità erano le petroliere che hanno causato la catastrofe ecologica vicino a Trinidad e Tobago nel 1979?
q_61 Quali esseri viventi sono in grado di assorbire l'anidride carbonica?
q_91 Qual era lo scopo della prima azione sostenuta da Greenpeace?
q_94 Qual è un fattore di rischio per le malattie cardiovascolari?
q_95 Quale è la categoria professionale più a rischio di cancro ai polmoni?
q_96 Dammi il nome di una parte dell'organismo attaccata dal virus Ebola.
q_98 Dammi un sintomo con cui si presenta l'affezione da virus Ebola.
q_101 Dammi il nome di un pesticida.
q_127 Quale animale tuba?
q_196 Qual è la valuta irachena?

An example of application of dynamic query strategy is CLEF2004question#14, *Come vengono chiamati i piloti suicidi giapponesi?*, for which we should be able to exploit a total match with a hypothetical lexical entry of the type (Fig. 52):

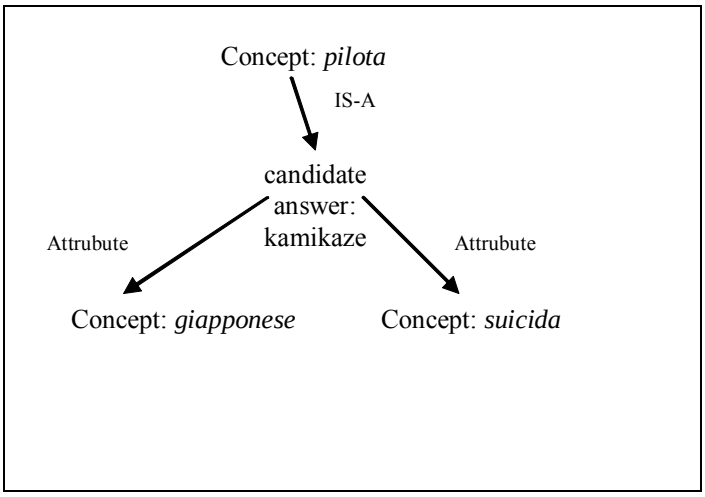


Fig. 52: a hypothetical lexical entry for *kamikaze* completely fulfilling the requirement of one of the question

Also a partial match, based in this case on the most selective link (the IS-A relation), may at least be used to propose a list of alternative candidate answers (in this case, all the hyponyms of the concept *pilota*, like for example the ones we can extract from IWN and illustrated in Fig. 53).

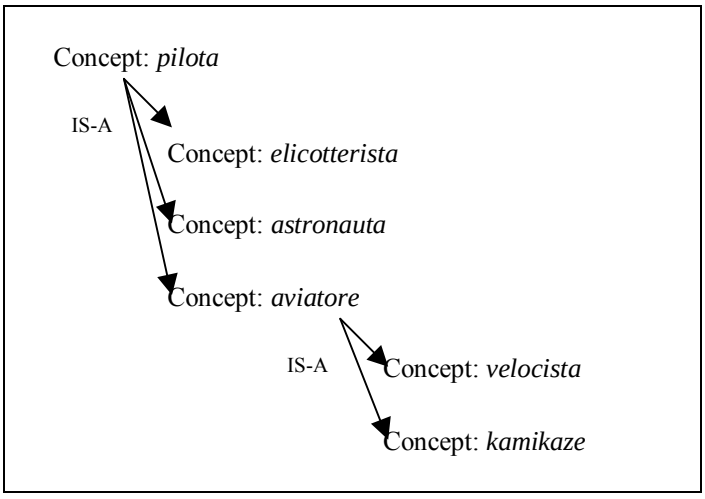


Fig. 53: hyponyms of the concept pilota as represented in IWN

The system should then be able to choose the "right" answer by exploiting the co-occurrence of the proposed answer and the other keywords of the question (in this example, the adjectives giapponese and suicida).

Only for some of these questions (question # 4, 14, 18, 24, 42, 61, 127, 196) the strategy gave the expected results. We will present the failed cases in chapter 5, where we will discuss the problem connected

to the exploitation of hyponymy in general (and where we will explain why the restriction of the use of dynamically created queries to the only cases of questions with form “Quale +tipo etc” can help to avoid the cases in which hyponymy is not the right way to individuate the answer).

However, what is really interesting is that this methodology assumes that the entire computational lexicon can be seen as a set of possible answers. Are there other semantic relations, beyond the ISA, that can be exploited to select the answer directly in the language resource? This is exactly what a knowledge base should be for: to provide a semantic representation useful for a specific reasoning task. We analysed the CLEF2004 test set and decided that in the IWN and SIMPLE-CLIPS there are at least three types of semantic relations that may be used to answer questions, i.e. the meronymy relation and the meronymy “made_of” relations in IWN and SIMPLE-CLIPS and the Derived_from relation in SIMPLE-CLIPS.

The meronymy/holonymy relation is instantiated as:

- meronym/holonym and their subrelations has_[mero|holo]_[part|portion|member] in IWN
- is_a_part_of/has_as_part and is_a_member_of/has_a_member_of relations in SIMPLE-CLIPS.

It can be exploited to answer questions like CLEF2004question#96: Dammi il nome di una parte dell'organismo attaccata dal virus Ebola (Name a part of the body that is affected by the Ebola virus) and TRECquestion#1059: What peninsula is Spain part of? (Di che penisola fa parte la Spagna?).

In order to successfully handle these cases, we expected the resources to encode the relation between the three possible answers (pelle, sangue, fegato) and the lexical entry organismo as well as the relation between Spagna and Penisola Iberica. Since in IWN two “parts” are encoded as separate lexical entries (parte anatomica/parte del corpo, parte del discorso) a preliminary check is done in order to exploit directly this hyperonym without analysis the form of the question to detect the part and the whole.

The Made_of relation (available as Mero_Made_Of in IWN and as Made_Of in SIMPLE-CLIPS) may nstead be useful when the question is of the type: “Di che cosa è fatto X?”. The AT of this type of question is SUBSTANCE but, before exploiting the hyponyms of the synset {sostanza, materia} and of the equivalent SemUs, the system creates a dynamic query with the subject of the question and with the lexical items (if any) that are in the LRs connected with the noun by means of a MadeOf type of link. If in the lexicon were a relation available like the one in Fig. 54, it would be possible to quite easily answer the CLEF2004question#115: Di cosa sono fatte le protesi mammarie?.:

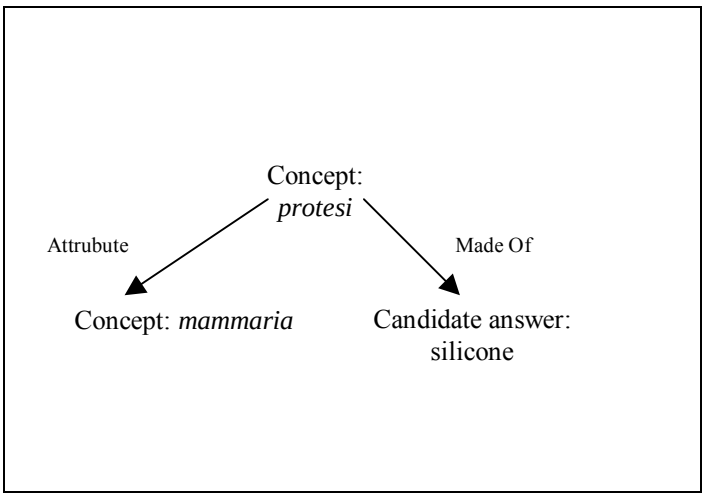


Fig. 54: a hypothetical lexical entry completely fulfilling the requirement of one of the question

In this case, the preliminary query connecting in AND the word protesi, mammaria, silicone would be much more selective and precise than the one that makes use of the entire SUBSTANCE taxonomy.

The relation Derived_from, available only in SIMPLE_CLIPS, may be useful to answer questions asking about the origin of concrete objects, like CLEF2004question#118: *da dove viene estratto l'acido salicilico?*.

4.3.4 Experiment on query expansion using IWN and SIMPLE-CLIPS

Query expansion is a technique consisting in automatically expanding the query by adding terms that are related to the words supplied by the user. Generally, the new terms are derived from lexical repositories, even if there are approaches that are based on the expansion by means of words statistically related, i.e. co-occurring with the original query keywords. We reported in chapter 2 all the systems among the participants to the TREC evaluation exercise that make use of a query expansion module. We now want to provide the enhanced prototype with a similar module that exploits synonyms and other semantically related terms available in LRs. As we said, without the possibility of making use of a sophisticated WSD module, the results reported here represent a hypothetical lower bound of the performance to be expected.

We decided to test the expansion only on the subset of questions for which the baseline system did not extract any pertinent (containing the answer) paragraphs.

An important reference for our work was the above said (see 2.5.2.2) (Magnini and Prevete, 2000), which reports on substantial improvements when using query expansion based on ItalWordNet synonyms[68].

[68] In their experiment, the identification of multiword expressions in the query and the disambiguation of the keywords were performed manually.

Moreover, the EWN model was analysed for Cross-Language Information Retrieval in (Gonzalo *et al*., 1998). Other experiments exploiting the American WordNet (among the others the one presented in Vorheers, 1994 and Mandala *et al*., 1998), present much less optimistic results.

In our experiment we expanded all the nouns, verbs and adjectives with relevance score >= 5 and <= 7[69].

The expansion is performed by exploiting:

- synonyms, i.e. variants of the IWN synsets and SemUs linked by synonymy relation in SIMPLE-CLIPS.
- Cross-pos synonyms, like the ones we can derive from the list of synsets grouped by the XPOS_SYNONYM relation in IWN and by the *EventVerb, DeverbalNounVerb, StateVerb, ProcessVerb* relations in SIMPLE-CLIPS. We hope, in this way, to provide the system with information that allows it to expand *corsa* (run) with *correre* (to run), *anticipo* and *anticipare* etc. With the same intent we extract the SemUs in SIMPLE-CLIPS that share the same predicate with a Master, VerbPastParciple or ProcessNominalization typeOfLink (in this way obtaining clusters of SemUs of the type *accusare, accusa* (to accuse, accusation) etc).
- Role relations: The Agent/patient_Role/involved relations from IWN were used, together withthe SIMPLE-CLIPS SemUs related to the predicate by means of AgentNominalization or PatientNominalization types of link and by means of the relations *AgentVerb*, *PatientVerb*. As far as IWN is concerned, we exploited only links not marked as reverse ("rev"), in order to avoid the generation of non valid inferences.
- adjectives pertaining to names of location and the locations themselves. The couple location-adjective is retrieved in the ItalWordNet database by extracting all the instances of the classes of type locations[70] and the adjectives that are linked to them by means of a HAS_PERTAINED relation. In this way, we obtain a list of couples of the type {America, Stati Uniti} – {americano, statunitense}, {Italia – italiano} - {Russia – russo} etc. The same type of information can be derived also from Simple-CLIPS, by exploiting the 270 concatenations of the relation *PolysemyNationality* (that link the adjective to the noun with identical meaning, like adj-*tedesco* and noun-*tedesco*) and *LivesIn* (that link the noun with the name of the country, like noun-*tedesco* and PN-*Germania*) (cf Fig. 55).
- IsAfollowerOf relation in SIMPLE-CLIPS (exploitable to extract couple of the type cattolico (N) – cattolicesimo).

It is worth remembering that in a similar experiment described in (Magnini and Prevete, 2000) the source for the second type of information, the one they described as morphological derivation, is not the computational lexicon but an Italian monolingual dictionary. We decided instead to try to exploit the

[69] Also the terms that expand the query are saved in the QuestionAnalysis xml file, in appropriate subelements of the morphological words.

[70] {paese, nazione, stato}, {continente}, {regione} and {città, urbe, centro urbano}

available semantic relations in both the LRs, since their linguistic design seems to guarantee the possibility of extracting this type of information[71].

Moreover, we decided to expand the query only at one level, i.e. by exploiting only the target of the relations having as its source the keywords of the query. Nevertheless, this choice has to be semantically declined: if in IWN all the information is structured around the notion of the *synset* in the attempt to preserve the cohesion of the concept, in SIMPLE-CLIPS the semantics of the concept are articulated in different SemUs. For this reason, when exploiting SIMPLE-CLIPS, we decided to add, to the SemUs of the first level of relation, their synonyms, in an attempt to create the same conditions we have when we exploit the IWN synsets. This means that in the case of CLEF2004question#45 *Quante interruzioni pubblicitarie durante i film sono attualmente permesse dalla CEE?* (How many commercial breaks during films are allowed by EEC at present?), the noun *interruzione* (interruption) will be expanded following its EventVerb relation (*interrompere*) and then with the synonyms of its target (*sospendere*):

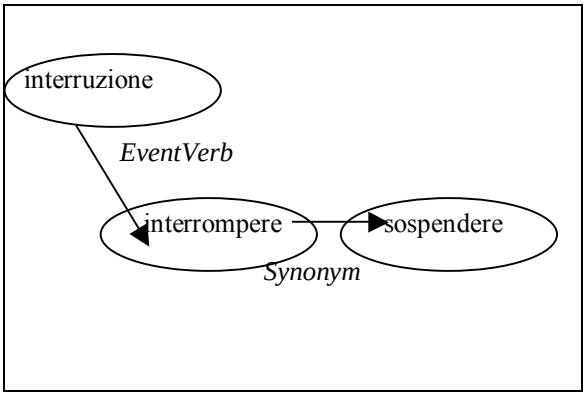


Fig. 55: the chain for the expansion of the SemU interruzione in SIMPLE-CLIPS

In the case of exploitation of the couple adjective-name of location of the type *tedesco-Germania* in SIMPLE-CLIPS, we will exploit a concatenation of relations but only to extract the two SemU at the edges of the chain. As a matter of fact, in SIMPLE-CLIPS the SemU linked to the name of the Location is always a noun, in turn linked to the adjective we found in the question. Therefore, the nominal SemU works as a bridge between the PN of the Location and the adjectives. In this case, however, we do not expand the query with the nominal SemU, but only with the PN and the adjective.

[71] One of the things that make SIMPLE-CLIPS different from IWN is the lack of explicitly bi-directional relations. This means that, for example, if we look in the lexical entry of *finire*, we will not find the link to the noun fine, that has to be sought instead in the lexical entry *fine*.

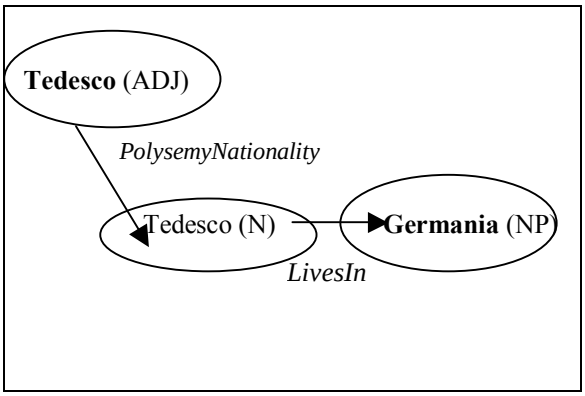


Fig. 56: the chain for the expansion of the adjective *tedesco* in SIMPLE-CLIPS

Such a strategy, if recursively applied, would give rise to the implementation of actual *lexical chains* of the type emerged in the analysis of the data of the questionnaire and described in (Harabagiu and Moldovan, 1998).

Important factors prevent us from exploiting rules recursively: first of all, the query expansion module is not the right place to implement recursive lexical chains, because the system does not "know" when/where to end the chain. In query expansion, the real danger is the explosion of the query, that would project the keywords semantically too far from the original query by a centrifugal force that would make the recall increase too greatly and the precision collapse. Differently, when trying to exploit lexical chains, we want to bridge the gap between two known/given texts.

Moreover, if we really want to implement lexical chains, we need formally and semantically valid heuristics capable of driving the dynamic discovery of meaningful paths among the thousands of possible connections of semantic relations (like the one proposed in Harabagiu and Moldovan, 1998 and transferred to the IWN and SIMPLE-CLIPS reality in Bertagna, 2004).

Nevertheless, even if the notion we are proposing is partially different from the one we found in the literature on lexical chains, we call every concatenation of semantic relations as the ones presented in Figures 11 and 10 a "*chain"*.

4.3.4.1 COMPOSITION OF THE QUERY WITH QUERY EXPANSION

Synonyms and other expansions of the basic terms are composed in Boolean expressions by using an OR connector. The complete sequence of steps can be described as follows:

Loop#1	Keyword(rel>2) connected by AND;
Loop#2	Keyword(rel>2) connected by AND and expanded;
Loop#3	Keyword (rel >5) connected by AND and expanded, Keyword (rel=5) expanded and connected by OR (if only one keyword with rel=5 is present, then loop#4)
Loop#4	Keyword (rel >5) connected by AND and expanded
Loop#5	Keyword (rel =10) connected by AND and expanded, Keyword (rel=7) expanded and connected by OR (if only one keyword with rel=7 is present, then

	loop#6)
Loop#6	Keyword (rel = 10) connected by AND and expanded

Table 14: query composition with query expansion

In order to illustrate the contribution of the lexical resources, we provide, in the next table, the question set of our experiment (i.e. questions for which the baseline prototype did not return any paragraph containing an answer), followed by the SemUs and the synsets that can be used to expand the query according to our specifications. The synsets and SemUs are in **bold** when the heuristics used for the Word Sense Disambiguation[72] fail to individuate of the correct sense.

[72] The "first sense in the resource" heuristic for SIMPLE-CLIPS and the "commonest sense in the corpus" heuristc for IWN.

question	Expanding terms from SIMPLE-CLIPS	Expanding terms from IWN
q_4: Qual è l'unità di misura di frequenza?		
q_10: Quando è stato consegnato il premio Nobel per la pace a Yasser Arafat?	consegnare(pred: consegna)	{consegnare} (Xpos_synonym: consegna)
q_12: A quanto ammonta il numero dei profughi palestinesi che si sono rifugiati in Libano?	palestinese(LivesIn-PolisemyNationality: Palestina), Libano (PolisemyNationality-LivesIn: libanese)	{profugo|rifugiato|fuoriuscito}
q_14: Come vengono chiamati i piloti suicidi giapponesi?	giapponese(PolisemyNationality-LivesIn: Giappone)	{giapponese}(pertains_to: Giappone)
q_17: A quale partito apparteneva Hitler?		{appartenere} (xpos: appartenenza)
q_18: Che lingua si parla in Germania?	Germania(LivesIn-PolisemyNationality: tedesco)	{**parlare**}{Germania|Deutchland}
q_28: Qual è il titolo del film di Stephen Frears con Glenn Close, John Malkovich e Michelle Pfeiffer?		{opera_cinematografica|film}
q_30: Chi fu il primo presidente degli Stati	Primo(Synonym: iniziale|principale|primario}	{America|Stati Uniti d'America|USA|Stati Uniti} (has_pertained: statunitense|americano)

Uniti?		
q_31: Qual è la professione di James Bond?		
q_32: Chi interpretava James Bond nei primi episodi della serie 007?	**interpretare**(pred: interpretamento,interpretatore)	**{interpretare}**(XPOS_Synonym: interpretazione), serie (xpos: seguente|venturo|successivo|prossimo}
q_33: Chi interpreta James Bond nell'ultimo film della serie 007?	**interpretare**(pred: interpretamento,interpretatore), ultimo (synonym: ultimo|finale)	**interpretare**(XPOS: interpretazione){film|opera_cinematografica} serie (xpos: seguente|venturo|successivo|prossimo}
q_35:A chi si chiede un mutuo?		{richiedere|chiedere}
"q_41: A quale età Michael Jackson ha cominciato a cantare nel gruppo dei ""Jackson Five""?"	**cantare**(pred: cantata, canto,cantante)	
q_44: Chi è l'inventore del televisore?	inventore(pred:inventare, invenzione)	{ideatore|inventore}(role_agent: ideare|immaginare|congegnare|concepire) {teleschermo|tivù|tv|televisione}
q_50: Qual è il quotidiano italiano più letto?	italiano (LivesIn-PolisemyNationality: Italia), **leggere**	leggere (xpos:lettura)
q_51: Quante persone soffrono di obesità negli Stati Uniti?		{patire|**soffrire}**(xpos:pena|dolore|dolore_fisico|male|sofferenza) {America|Stati Uniti d'America|USA|Stati Uniti} (has_pertained: statunitense|americano)
q_52: Qual è l'ingrediente base della cucina giapponese?	giapponese(LivesIn-PolisemyNationality: Giappone)	**ingrediente**
q_57: Quando venne introdotto l'alfabeto	**introdurre**(pred:introducimento|introduttore)	**introdurre**(xpos:introduzione)

cirillico?		
q_62: Da che cosa è ricoperto il continente antartico?	**ricoprire**	**ricoprire**(xpos:ricopertura)
q_66: Quanti anni ha il papa?		
q_68: Quanti sono i cattolici nel mondo?	cattolico(IsAFollowerOf cattolicesimo), **mondo**	{macrocosmo|mondo|cosmo|creato|natura}(near_synonym:universo)
q_71: Quanti stati in America hanno la pena di morte?	**Pena**, morte(eventVerb:morire e pred: morto,morire)	**stato**{pena_di_morte|morte|pena_capitale}{Americhe|America}(1)(has_pertained:americano)
q_72:Quante esecuzioni capitali ci sono state negli Stati Uniti nel 1993?	esecuzione (EventVerb: eseguire)	**esecuzione**{America|Stati Uniti d'America|USA|Stati Uniti} (has_pertained: statunitense|americano)
q_73: In che anno è stato abolito l'apartheid in Sudafrica?	abolire (pred:abolizione)	**abolire**
q_74: A quanto ammonta la popolazione degli USA?		{popolazione|cittadinanza}{America|Stati Uniti d'America|USA|Stati Uniti} (has_pertained: statunitense|americano)
q_82: A quale età i giovani vengono convocati per sottoporsi alla visita di leva?	convocare (pred:convocato,convocazione,convocatrice), visita (DeverbalNounVerb: visitare, Pred: visitamento), **leva**	
q_86: Quale ditta è accusata di avere sfruttato il lavoro minorile?	accusare(pred: accusa, accusato, accusatore), lavoro (DeverbalNounVerb: lavorare, pred: lavoratore, lavorante, lavorare, lavorazione, lavoratura)	{tacciare|incolpare|accusare|imputare} (Xpos: denuncia|accusa)
q_89: Dove	**fondare** (Pred: fondato)	{fondare} (xpos_syn: fondazione)

venne fondata Greenpeace?		
q_91: Qual era lo scopo della prima azione sostenuta da Greenpeace?	scopo (Synonym: intento, obiettivo,meta,mira,target,funzione), azione (Synonym: agire)	**azione**(asseverare|sostenere|dare_per_certo|asserire}(senso sbagliato)(xpos: dichiarazione|affermazione|asserzione)
q_95: Quale è la categoria professionale più a rischio di cancro ai polmoni?	{categoria|tropo}(1), rischio (pred:rischiare),cancro (Synonym: tumore)	**Rischio**, cancro(near_synonym:neoplasma|neoplasia|tumore)
q_96: Dammi il nome di una parte dell'organismo attaccata dal virus Ebola.	**attaccare** (Syn: aderire)	{**attaccare**} (xpos:attaccatura)
q_98: Dammi un sintomo con cui si presenta l'affezione da virus Ebola.	**presentare**(pred: presentazione, presentatore)	affezione (near_syn: malessere|disturbo) involeved_patient: ammalato|malato)
q_114: Quando si sposarono il principe Carlo e Diana?	sposare (pred: sposato, sposamento, sposo, sposa)	
q_120: Come può venire trattata un'allergia?	**trattare**(pred: trattatrice,trattatore)	{**trattare**)
q_124: Che cos'è la massoneria?		
q_127: Quale animale tuba?		
q_128: Quanti panda ci sono allo stato brado in China?		
q_143: In quale anno, prima del 1995, si è tenuta la Conferenza		{femmina|donna}mondiale(pertains_to macrocosmo|mondo|cosmo|creato|natura)

mondiale sulle donne?		
q_147: Quante persone vivono a Bombay?		
q_148: Chi è il direttore della IAEA?		
q_190: Quanto è alto il K2?	**alto**	**alto**|elevato
q_192: Quale paese è il campione del mondo di calcio?	**campione** (Synonym: saggio)	{macrocosmo|mondo|cosmo|creato|natura}{football|calcio}

Table 15: expanding the query using the IWN synsets and SIMPLE-CLIPS SemUs

4.3.4.2 RESULTS OF THE QUERY EXPANSION EXPERIMENT.

The results of the experiment show that only a small improvement is obtained by expanding the query. On 43 "non-answered" questions, there are respectively four and one cases of retrieval of new and pertinent paragraphs when using IWN and SIMPLE-CLIPS. The average precision increases only by few points going from 0 to 0.005 when exploiting IWN and to 0.002 when using SIMPLE-CLIPS.

The rest of the questions remain without an answer.

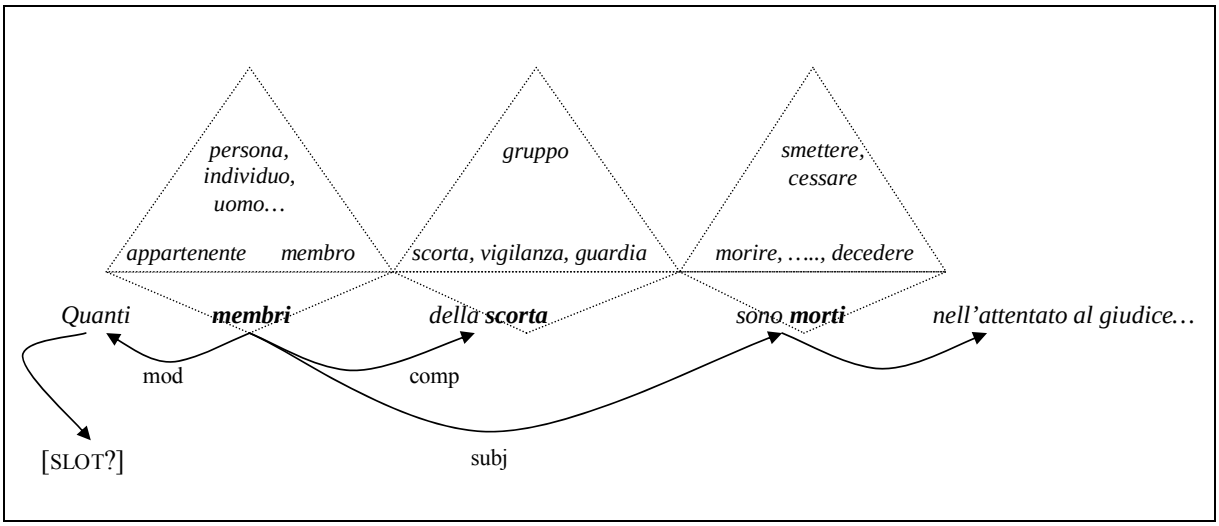
The results will be discussed more in depth in 5.4.9 and 5.4.10.

4.3.5 Answer Detection and Extraction

The strategies for answer individuation and extraction of the enhanced prototype are almost the same as those adopted in the baseline version. But, when the mere exploitation of syntactic patterns are not enough to individuate the answer, LRs intervene to empower the rules described in 4.3.2.7.1 by lexically expanding the cases to which the rules apply.

Examples of the strategy involving LRs are CLEF2004question#7 (*Quanti membri della scorta sono morti nell'attentato al giudice Falcone?*[73]) and CLEF2004question#64 (*Cosa può causare il tumore ai polmoni?*[74]).

Fig. 57 provides a graphical description of the type of analysis foreseen. We have chosen to represent the lexical-semantic explosion of the arguments and predicates of the question as triangles having as a base the synonyms of the query term and as a summit angle their hyperonymic concept.



[73] *How many members of the escort died in the attack to Judge Falcone?*
[74] *What causes lungs tumor?*

Fig. 57: example of lexical/semantic layer of representation

The search strategy is thus augmented by adding to the terms involved in the iteration described in 4.3.2.7.1 the terms in the triangles:

mod([slot?], (membro|appartenente|persona|individuo|uomo|essere_umano))
comp((membro|appartenente|persona|individuo|uomo|essere_umano), (scorta|vigilanza|guardia)
subj((membro|appartenente|persona|individuo|uomo|essere_umano), (morire|decedere)
comp (morire, attentato) etc..

In this case, the exploitation of the IWN IS-A relation between the word *membro* (*member*) and *uomo* (*men*) helps to individuate the answer in the retrieved paragraph:

"*..nella strage di Capaci… dove furono uccisi il giudice Giovanni Falcone ..e tre uomini della scorta..*"[75].

In the second case, the synonymy between *causare* (*to cause*) and *provocare* (*to prove*) on one hand and *tumore* (*tumor*) and *cancro* (*cancer*) on the other helps to match question to the candidate answer text:

"*…alimentando l'ipotesi…che gli scarichi diesel provochino il cancro*"[76].

Moreover, LRs can be used to confirm the answer found by pattern matching on the functional relations. For example, in answering CLEF2004question#11 (*Quale è la città sacra agli ebrei?*), LRs are used to confirm the correctness of the answer provided by the system by evaluating its semantic type: in this case, *Gerusalemme* is classified in IWN as a city so the type is the same as the expected answer. This second strategy does not add anything to the final performance of the system but contributes to providing more accurate answers.

LRs are also used to generate the answer in the case of paragraphs obtained via dynamic querying. In the case of CLEF2004question#4 *Qual è l'unità di misura della frequenza?*, the most successful retrieval is the one the system obtained when it submitted to IXE the query "*Herzt* AND *frequenza*", where Hertz is one of the entries that, both in IWN and SIMPLE-CLIPS, are classified as "*unità di misura*". In this way, the lexical entry itself is provided as the answer.

[75] *..in the Capaci massacre…where Judge Falcone..and three men of his escort died..*
[76] *..it fosters the hypothesis that…diesel exhaust provokes cancer*

4.3.6 Schematic description of the extraction strategies

In this paragraph we provide a summary of the different extraction strategies adopted in the enhanced prototype. A number of rules are applied on the basis of the information gathered during the Question Analysis phase, i.e. the Question Stem and the Answer Type Term. Nevertheless, the most important information for the definition of the extraction strategies is the Answer Type. Fig. 58 shows the various types of information that contribute to the strategies adopted in the extraction module.

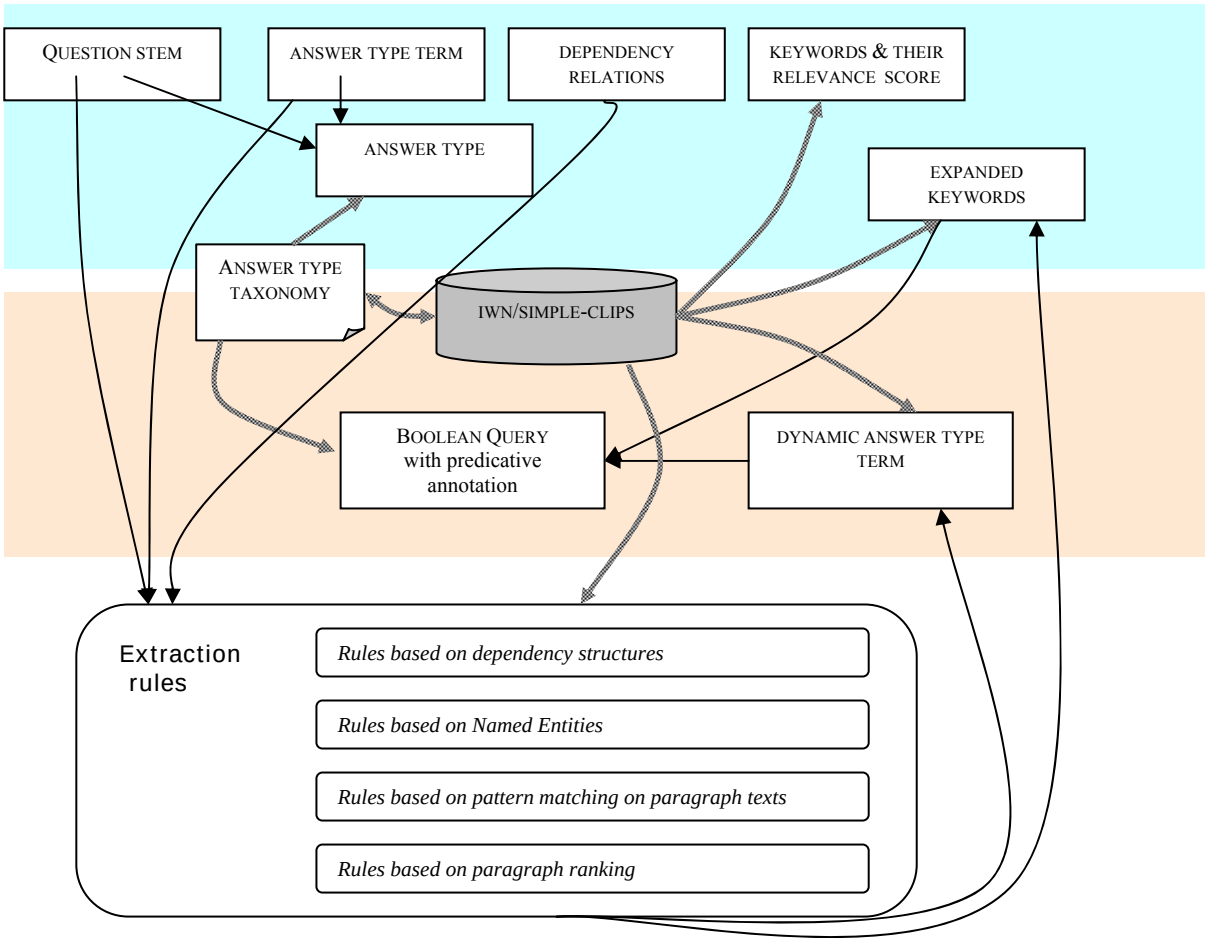


Fig. 58: information exploited in the answer extraction module

Three fundamental cases can be identified:

Case#1: the Answer Type is DEFINITION.

Two situations are tested:

- if Question_Stem=[Che cosa|Che] (*Che cosa è il Mossad?* What is the Mossad?), the strategy based on pattern matching on the paragraph text is applied (*Il Mossad (il servizio segreto israeliano)*…, The Mossad (the Israeli secret service)…) .

- if Question_Stem=[Chi] (*Chi è Silvio Berlusconi?* Who is Silvio Berlusconi?), the rule based on the presence of a Dependency Relation of type "adposition" is applied (*..il Presidente del Consiglio Silvio Berlusconi,* the Prime Minister Silvio Berlusconi…) . If the answer is not found, then a rule based on the presence of a relation of coordination is applied (*Silvio Berlusconi, presidente del consiglio*, Silvio Berlusconi, Prime Minister…) followed by a control on the semantic type of the candidate answer (that has to be of type Human).

Case#2: the Answer Type value is different from DEFINITION and corresponds to a Named Entity category (*Quale città ..?*, What city…?, *Quale pittore..?*, What painter…?, *Quanto è alto…?*, How tall is…?).
The system has to choose among the paragraphs extracted by the Search Engine by restricting the search to the correspondent NE. On these paragraphs, the search of the answer based on dependency relations is applied and, if not answer is found, the search is iterated by expanding the nodes of the dependency representation with synonyms and hyperonyms. The answer has to be of the right type (it has to correspond to the expected Name Entity category). If also this strategy fails, the system selects as answer the Named Entity of the paragraph with the highest ranking.

Case#3: the Answer Type is different from DEFINITION and does not correspond to a Named Entity category (*Come si chiamano i piloti suicidi giapponesi? Qual è l'unità di misura della frequenza?* etc.).
The strategy based on exploitation of the dependency relations is applied (also by expanding the rules with synonyms and hyperonyms). If no answer is found, then new dynamic queries are submitted to the Search Engine, by substituting the Answer Type Term with its hyponyms. If a single iteration gives results, then the system provides the hyponym as answer to the question. If more than one hyponym give positive result, the "right" answer is identified by using a syntactic criterion.
In case#3, some ad-hoc rules are also envisaged, as the one tailored to answer questions with ATT *professione* (*Quale è la professione di..?*, What is the profession of…?) and questions for which exploitation of the *meronymy* and of the *Derived_from* relations is foreseen.

4.3.7 Moving towards inferential chains: is it feasible?

A possible follow-up of our work would be the creation and exploitation of something similar to the inferential chains described in par. 2.5.2.4. We adopted the same methodology of (Harabagiu *et al*. 1998) to discover significant inferential paths through the large set of semantic relations of ItalWordNet and through the rich connectivity (ranging from the argument structure to the qualia roles) of CLIPS. This is something we already did when we created the paths for the QA pairs of the questionnaire. One of the things that distinguish our experiment from Harabagiu *et al*.'s work is that the types of information that in (Harabagiu *et al*., 1998) are derived from the WordNet glosses are supposed to be already available in EuroWordNet and CLIPS (where relations between different POS are allowed and envisaged). Enabling the recognition of

inferential paths could play an important role in filling the gap between the question and the answer, as it is evident in the following example: Q: *Quale funzione ha la milza?* (Which is the function of the spleen?) A: La milza produce linfociti (The spleen products lymphocytes). In this case, since no direct relation is established between *funzione* (function) and *produrre* (to product), in order to expand the query with potential relevant terms, the system should be able to resort to a complex heuristic. In this case the significant path through ItalWordNet would be:

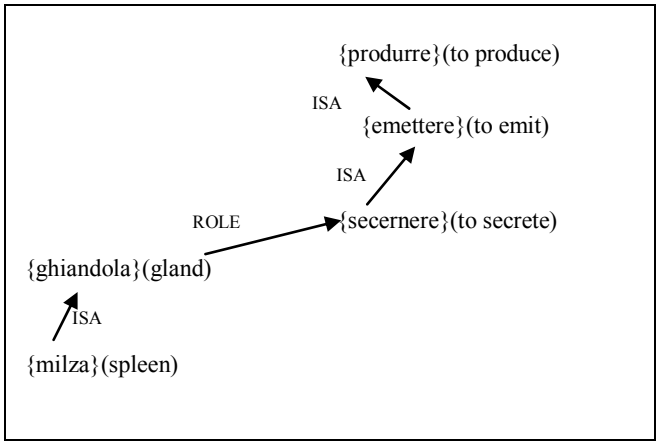


Fig. 59: concatenation of semantic relations connecting *milza* and *produrre* in IWN

In this example, the ISA inheritance mechanism triggers the inferential rules which allow us to derive:

milza ISA ghiandola+ ghiandola ROLE secernere= milza ROLE secernere
milza ROLE secernere+secernere ISA emettere=milza ROLE emettere
milza ROLE emettere+emettere ISA produrre= milza ROLE produrre

The primitive rules are thus:
c1 ISA c2+ c2 ROLE c3= c1 ROLE c3
c1 ROLE c2+c2 ISA c3=c1 ROLE c3

Starting from the complete list of the almost 75 EWN semantic relations, we have studied all the possible relation pairs. Not all the available relations can be combined to generate valid primitives since some relations can be applied only to specific POSs (it is not possible to combine, in this order, a ROLE relation, which applies between nouns or between a noun and a verb, with a MANNER_OF relation, which goes from an adverb to a noun or a verb). By avoiding combinations not respecting the right POS concatenation, we obtained 603 relation pairs. Moreover, the fundamental EWN distinction between first, second and third order entities prevents us from pairing relations whose concatenation doesn't respect correct

entity order (in this sense a HAS_HOLONYM, which applies between first order entities, and an INVOLVED, that links a second order entity to a first order entity, cannot be combined). We found about 80 cases of this type. At the end, about 480 formally valid relation pairs was formed and evaluated. When having to choose a name for the result of the concatenation we preserved, if possible, the name of "normal" EWN relations. This allows us to more easily create complex *inferential rules* resulting in further concatenations of relation pairs. Moreover, we preferred to eliminate, in the resulting name, any indication of the cross-parts of speech nature of the relation. This because the primitive rules are supposed to represent a totally semantic link between not adjacent concepts and any reference to morphosyntactic features of the relation is not meaningful.

We can discover endless possible ways to navigate along the relations in IWN, but the key is to find only fundamental concatenations that support inference. We tried, then, to verify if the paths based on this large set of primitive rules can be of any help in the QA task. We have to specify that we haven't implemented yet an automatic procedure to extract the resulting semantic paths: we have worked manually on question-answer pairs of the CLEF QA campaign, extracted using IXE. Unfortunately, results are not encouraging: only a very small number of questions can be answered expanding the query with concepts belonging to semantic paths driven by the inferential rules. Potentially, the linguistic design seems suited to support text inference but the number of available links and connections is too low to be useful on an extended, open-domain task. An example is question_#4: *Quando e' stato stipulato il Trattato di Maastricht?* (When was the Maastricht Treatry draw up?). The three keywords (*Trattato and Maastricht* and *stipulare*) are not enough to retrieve any passage, while with only (*Trattato and Maastricht*) we obtain a high recall of about 300 paragraphs. But how can the system pinpoint the "answer" among this large set paragraphs? The presence in the paragraph of a named entity of the type "Date" is not enough to discriminate (since in almost all the paragraphs there is at least one temporal expression). We found 4 possible candidates:

"…ratifica del Trattato di Maastricht…vinto…. nell'autunno del 1992" (ratification of the Maastricht Treaty …won…autumn 1992)

"…conclusione del Trattato di Maastricht nel 1991" (conclusion of the Maastricht Treaty in 1991)

"..secondo referendum di ratifica dopo quello..del settembre '92….del Trattato di Maastricht" (the second referendum after the one in September '92)

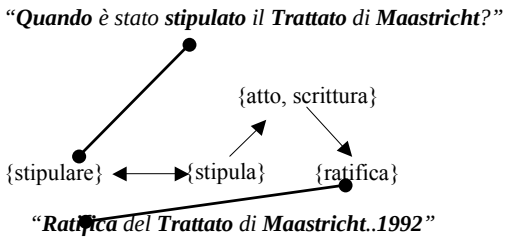


Fig. 60: semantic path between *stipulare* and *ratifica* in IWN

the inferential path is traced by the primitives:

stipulare XPOS_NEAR_SYNONYM *stipula*+ *stipula* HAS_HYPERONYM *atto*=
stipulare HAS_HYPERONYM *atto*
stipulare HAS_HYPERONYM *atto*+*atto* HAS_HYPONYM *ratifica*= *stipulare* CO_HYPONYM *ratifica*

Altough a number of primitive rules (consisting of a pair of relations) can be identified, finding regularities in the way significant paths can be identified in the resources is very difficult. What is interesting to note is that one of the main shortcomings of semantic language resources (as far as reasoning in concerned) seems to be the richness of expressive modalities. The very large set of semantic relations (e.g. in ItalWordNet more than 70 types of relation are available) allows the lexicographer to encode differently information of very similar types, increasing the computational complexity in the discovery of the useful paths (and increasing also the possibility to make mistakes choosing not appropriate relations). In general, it seems that for the task at hand, it would be better to keep low the number of relation types while increasing the number of connections of the same types. Moreover, although the richness of the expressive modalities, the information is not consistently distributed (many relations are very rare). This is not true for SIMPLE, where the use of Templates plays an important role in making the distribution of information types more consistent and equilibrate. In the paper we will illustrate the problematic cases. In general, exploiting LRs to support complex inferences seems a very hard task, in particular for the computational heaviness of the required elaboration.

4.3.8 Exploiting the system output to enrich the lexicon

In paragraph 4.3.8 we introduced the idea we have of the interplay between information residing in semantic lexicons and applications. We define the general model we have in mind as a *closed*, integrated framework, where the application exploits the content of the semantic resources but, at the same time, provides a feedback that can enrich the content of the lexicon itself. When the system answers a question it provides information that can be of three types:

- information about facts that should not be listed in a semantic lexicon (what we may call aencyclopedic information). This is the case, for example, of the answer to question *Quanti membri della scorta sono morti nell'attentato al giudice Falcone?* ("3"), or of question *In quale anno venne conferito il Nobel a Thomas Mann?* ("1929") etc. In these cases, we do not want our lexicon to host the complete factual information regarding the escort to Judge Falcone or the year of assignment of the Nobel Prize to Thomas Mann.
- lexical semantic information already expressed in the lexicon as such (kamikaze as hyponym of pilota, Hertz as hyponym of unità di musira etc.).

- lexical semantic information not yet available in the lexicon (*Quale città è alla confluenza del Reno e della Mosella? Coblenza. Quale ingrediente è alla base delle cucina giapponese? Tofu*). Also in this case we do not want to save the aenciclopedic information regarding the meeting of the Rhine with the Mosel but it would be quite interesting to acquire the new entry (Auto_SemU of monovariant Auto_synset) *Coblenza* and classified it under the node {città} already available in the lexicon. Not only a new entry may be inserted in IWN and SIMPLE-CLIPS, but also a new semantic link. If the system is able to answer the question *Quale ingrediente è alla base delle cucina giapponese?* (*tofu*) by means of pattern matching on syntactic dependency structures, an Auto_Synset_relation may be created between the already encoded synset {tofu} (classified in IWN as a cheese) and the synset {ingrediente}.

Within this dissertation, no actual implementation of the mechanism of feedback has been realized. However, the exploitation of part of the impressive amount of implicit semantic information available in not structured texts is a potential way to enrich the static, fixed content of lexical repositories like ItalWordNet or SIMPLE-CLIPS.

Open-Domain QA is an application whose ouput is generated by analysing free text and by extracting from it relevant information. In doing so, QA not only exploits lexicons but can enrich them with new information. Obviously, this is something that may be interesting only in the case of a very robust application that is able to efficaciously handle hundreds of questions and process very large corpus data (requirements that our system, for its prototypical nature itself, does not have). The idea is that the answers extracted by exploiting the dependency analysis or the recognition of Named Entity in the text can be stored in the lexicon and structured according to the classifications already available in the lexicon. The final aim is obviously the reusability of the output of the application, in the same application or in others.

What is interesting and would deserve an in-depth study is the analysis of the difference between the already encoded information and the one that would be acquired from the corpus. Some information would be of the same type: the potential new synset {Valentina Terechkova}, instance of the synset {cosmonuta}, is surely of the same type of the already encoded synset {Leon Battista Alberti}, instance of {architetto}. But what is the difference between the "logically consistent" hyperonymy between {tofu} and {formaggio} and the acquired auto_relation between {tofu} and {ingrediente}? We think that the acquisition of these new information from corpora would be however extremely interesting from the point of view of equipping language resources with a more *fuzzy* and context-based type of information.

A comparison between the results obtained by the Enhanced prototype and the performance of the baseline prototype will be provided in Chapter 5.

5

In this final chapter, we analyse the results obtained by the prototype on the CLEF2004 test bench, highlighting both successful exploitation of the information stored in language resources and the problems encountered.

5.1 Comparative results

The following two tables present the comparison between the performance of the two versions of the prototype (on the right side in each column are the baseline results, on the left side are the enhanced results). Results are given in separate tables respectively for the enhancement obtained by exploiting ItalWordNet and SIMPLE-CLIPS.

#Answer	#Right	#Wrong	#IneXact	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy	
							Precision	Recall
200	91-111	87-71	22-18	45.5-55.5	42.7-53.3	70-75	0.62	0.5

Table 16: comparison between the results of the baseline and of the IWN-based enhanced prototypes

#Answer	#Right	#Wrong	#IneXact	Overall Accuracy %	Accuracy over F %	Accuracy over D %	NIL Accuracy	
							Precision	Recall
200	91-100	87-81	22-19	45.5-50	42.7-47.2	70-75	0.62	0.5

Table 17: comparison between the results of the baseline and of the SIMPLE-CLIPS-based enhanced prototypes

Table 18 and Table 19 show the comparison between the "wrong results" of the two prototypes, classified according to their question stem (results are given respectively for the exploitation of IWN and SIMPLE-CLIPS).

Question Type	# of questions in the test set	% Wrong	Improvement (percentage points)
Quale (pronoun)	17	76.4-58.8	17.6
Come si chiama	6	33.3-16.6	16.7
Quale/ Che (adj)	43	46.5-34.8	11.7
(Che) Cosa (pn)	14	25-14.2	10.8
Others (dimmi, dammi, nomina)	7	57-42.8	14.2
Quanto (pn)	9	55.5-44.4	11
Quanto (adj)	18	55.5-50	5.5
Chi	35	34.2-31.4	2.8
Quanto (adv)	1	100	0
Come	12	58.3	0
Dove	14	35.7	0
Cosa (DEF)	10	33.3	0
Quando	14	21.4	0

Table 18:not answered questions classified according to their question stem (by using IWN)

Question Type	# of questions in the test set	% Wrong	Improvement (percentage points)
Quale (pronoun)	17	76.4-64.7	11.7
Come si chiama	6	33.3-33.3	0
Quale/ Che (adj)	43	46.5-39.5	7
(Che) Cosa (pn)	14	25-14.2	10.8
Others (dimmi, dammi, nomina)	7	57-42.8	14.2
Quanto (pn)	9	55.5-44.4	11
Quanto (adj)	18	55.5-55.5	0
Chi	35	34.2-34.2	0
Quanto (adv)	1	100	0
Come	12	58.3	0
Dove	14	35.7	0
Cosa (DEF)	10	33.3	0
Quando	14	21.4	0

Table 19: not answered questions classified according to their question stem (by using SIMPLE-CLIPS)

In the next two tables, we provide an overview showing the difference between the performances of the enhanced prototype when it exploits IWN and SIMPLE-CLIPS. The comparison is made on the general accuracy of the system (Table 20) and on the specific types of question successfully analysed by the system (table 6).

Overall Accuracy using IWN %	Overall Accuracy using SIMPLE-CLIPS %	Difference in the overall accuracy (percentage points) between the baseline and the IWN-based enhanced prototype.	Difference in the overall accuracy (percentage points) between the baseline and the SIMPLE-CLIPS-based enhanced prototype.
45.5-55.5	45.5-50	10	4.5

Table 20: difference in the overall accuracy obtained by exploiting the two lexicons

Question Type	Improvement (percentage points) using IWN	Improvement (percentage points) using SIMPLE-CLIPS	Difference in the obtained improvement (percentage points)
Quale (pronoun)	17.6	11.7	5.9
Come si chiama	16.7	0	16.7
Quale/ Che (adj)	11.7	7	4.7
(Che) Cosa (pn)	10.8	10.8	0
Others (dimmi, dammi, nomina)	14.2	14.2	0
Quanto (pn)	11	11	0
Quanto (adj)	5.5	0	5.5
Chi	2.8	0	2.8
Quanto (adv)	0	0	0
Come	0	0	0
Dove	0	0	0
Cosa (DEF)	0	0	0
Quando	0	0	0

Table 21: comparison between the improvement obtained with IWN and the one obtained with SIMPLE-CLIPS

The comparison shows that, even if the results are quite similar, some significant differences can be detected when using the two resources.

Results shown in Tables 1, 2, 3 and 4 are interesting: it is possible to observe an overall improvement determined by the exploitation of the two LRs. The improvement is obvious when one considers the ten percentage points that divide the two prototypes in general accuracy but also when we consider its distribution on the various types of question.

The types of question whose results improved in the most evident way are the ones we thought would have taken more advantage from LR exploitation (i.e. the ones for which the system has to analyse the ATT in order to individuate the expected answer type): questions introduced by *Quale* (both in adjectival and pronominal function), but also by the various imperatives *dammi* (give me), *dimmi* (tell me), *nomina* (name) and by the frequent interrogative form *"come si chiama…?"* (What's the name of..?).

However, these types of question are always the ones that have the highest degree of system failure and it is not easy to formalize strategies to handle the half of the questions that do not receive an answer.

In the following paragraphs we will analyse the reasons behind the system failures; this time we will not take into consideration, differently from what we did when we analysed the baseline results, failures deriving from the erroneous treatment of syntactic or morpho-syntactic information. We will try to organize this qualitative analysis on the basis of phenomena more directly connected to the methods adopted in the two lexicons to individuate and characterize the *conceptual/semantic content of the lexical item*. As already mentioned in the Introduction, these methods concern the following intertwined issues:

v) *granularity* of the representation of the ambiguity, i.e. the number of senses that is supposed to be appropriate;

vi) *breadth* of the lexicon, i.e. the number and type of lexemes admitted in the language resource

vii) *depth* of the lexicon, i.e. number and type of the linguistic phenomena described in the lexical entry and their usefulness in supporting reasoning and inference, with particular attention to aspects involving *connectivity* (the expression of relations with other elements of the lexicon).

Problems connected to these aspects are somehow transversal to all the modules of LR exploitation so we interpret them as structural problems, having general significance.

5.2 Granularity Issues

Problems connected to sense distinction arise in every single interaction between lexicons and application, not only when lexicons propose more than one sense for a lemma but also when a single sense is proposed, since it might be the "wrong" one. We are aware that in the system no actual WSD module is exploited: the "first sense in the corpus" heuristic is only a baseline and in this sense we can see the obtained performance as a lower bound that "can only get better". Nevertheless, no perfect WSD system exists at the moment and the problem of identification of the "right" sense of 100% of occurrences seems nowadays almost irreversible. The first module whose performance is negatively impacted by incorrect sense selection is the Answer Type determination. In 4.3.3.4 we described the methodology for enriching the Answer Type Taxonomy of the baseline prototype with a new layer of lexical-semantic information. Both resources allow the system to increase the number of identified expected answer types from the 126 of the baseline prototype to i) the 171 recognized thanks to IWN and ii) the 166 recognized by exploiting SIMPLE-CLIPS.

Nevertheless, some ATs were incorrectly identified and there are still about 30 questions for which the system was not able to derive any Answer Type.

Table 3 gives an overview of the improvement determined by the exploitation of LRs with respect to the results obtained with the baseline prototype. We can see that, together with other important factors that we will discuss in the next paragraphs, a reason behind failure in AT identification is the incorrect selection of

the word sense. In fact, while for IWN the "commonest sense in the corpus" heuristic is an almost valid aid for disambiguation, more cases of incorrect sense attribution are registered when using SIMPLE-CLIPS. Next table provides an overview of the number of ATs identified by the prototypes, together with the number of ATs identified in an incorrect way. As far as the enhanced prototype is concerned, the incorrect ATs are classified on the basis of the reason behind their incorrectness, by distinguishing between cases due to wrong sense selection and other reasons.

<table>
<tr><th></th><th># identified ATs on 200 questions</th><th colspan="2">#incorrect ATs</th></tr>
<tr><td>Baseline prototype</td><td>126</td><td colspan="2">2</td></tr>
<tr><td rowspan="3">Enhanced prototype (IWN)</td><td rowspan="3">171</td><td colspan="2">4</td></tr>
<tr><td>Incorrect WSD</td><td>Others</td></tr>
<tr><td>2</td><td>2</td></tr>
<tr><td rowspan="3">Enhanced prototype (SIMPLE-CLIPS)</td><td rowspan="3">166</td><td colspan="2">10</td></tr>
<tr><td>Incorrect WSD</td><td>Others</td></tr>
<tr><td>8</td><td>2</td></tr>
</table>

Table 22: identified ATs in the two versions of the prototype and results for the two LRs

Incorrect selection of the sense of *casa* (house) is for example at the base of the failure in AT identification for the two questions:

CLEF2004question#27: *Quale casa automobilistica produce il "Maggiolone"?* (What car company produces the "Beetle"?)

CLEF2004question#43: *Come si chiama la casa discografica di Michael Jackson?* (What's the name of Michael Jackson's record company?)

The selected sense, in both lexicons, was the one of *casa* as building, thus the derived AT is BUILDING>LOCATION in both cases.

Globally speaking, in SIMPLE-CLIPS the right sense of the ATT was missed 12 times. Not in every case, however, this has an effect on the AT identification. For example, in the case of CLEF2004question#113 *Come si chiama la compagnia di bandiera tedesca?* (What is the official German airline called?) the selected sense of *compagnia* (company) is not the commercial one but the one referring to *an informal gathering of people*. The two cases, however, share the same AT HUMAN GROUP, so the final result is not affected by the erroneous sense attribution.

A reflection on the nature of the distinctions that drive the sense splitting in semantic lexicons is needed: it seems that, for the majority of the sub-tasks encountered in our application, a coarse granularity in the definition and representation of the lexical items is sufficient to achieve good results. QA is somehow

"Named-Entity-Sensitive": each distinction that the system is able to capture at question analysis level has afterward to be appreciated during the answer detection phase. It means that being able to understand that the expected answer is the name of a ship does not have any positive consequences unless the system is also able to individuate the Named Entity class "ships" in the candidate answer. This surely has an effect on the granularity of lexical description that is required by this type of application and this can be observed when we evaluate the connection between the AT Taxonomy and the nodes of the lexical resources: in order to guarantee a successful recognition of the ATT and of other meaningful words of the question, we had to link some ATs to more than one sense of the same word. This happens, for example, for the Answer Type YEAR>DATE, that we decided to link to all the synsets in IWN with variant "*anno*":

{Anno} – *tempo necessario alla Terra per compiere il suo giro intorno al Sole* (the time employed by the Earth to turn around the Sun)

{Anno} – *periodo di dodici mesi in genere* (a generic period of twelve months)

{Anno} – *periodo di tempo non determinato, di cui si sottoline al lunghezza* (an undetermined period of time, usually very long)

{Anno} – *arco di tempo durante il quale di svolge un'attività* (a period of time, e.g. in agriculture; the span of an activity cicle)

The sense inventory of *anno* proposed by IWN and supported by valid Italian monolingual dictionaries (Garzanti, 2005, De Mauro, 2000, DISC, 1996) was already noted and discussed in (Calzolari *et al*., 2003). It is surely possible to roughly organize the occurrences of *anno* in the corpus according to the senses available in this inventory: for example (Garzanti, 2005) proposes the following distribution:

Anno 1: *anno siderale* (sidereal year), *anno astrale* (astral year), *anno luce* (light year), *anno civile* (calendar year).
Anno 2: *anno 1265* (year 1265), *anno prossimo* (next year), *anno nuovo* (new year), *quest'anno* (current year), *l'altr'anno* (last year), *gli anni Venti* (the twenties) etc.
Anno 3: *è un anno che aspetto l'autobus!* (I have been waiting for the bus for ages)
Anno 4: *anno scolastico* (school year), *anno accademico* (academic year), *anno liturgico* (liturgical year).

We need to know whether this kind of distinction can be captured by a computer program that analyses real text and, above all, if the distinction is really indispensable for an NLP task.
For a human being, the glosses are probably self explicative and it is not so difficult to catch the semantic difference between, for example, sense 1 and 2. One difference may be the more "astronomic" feel that words like *siderale* (sidereal), *astrale* (astral) and *luce* (light) have and by the referral to *Terra* (Earth) and *Sole* (Sun) in the definition (no explicit domain is indicated in the lexical entry).
Word Sense Disambiguation is the mapping between a textual occurrence and a sense in the lexicon, i.e. the problem of determining in which sense a word having a number of distinct senses is used in a given

sentence. But the attempt to select the "right" sense is bound to fail if there is no clear idea of what a *sense* is: word meaning seems to be a kind of *Holy Graal* and the *checklist theory* of meaning itself is suspect (Fillmore, 1975), with corpus evidences revealing loose and overlapping categories of meaning and standard meaning of words extended and contracted in a variety of ways (Kilgarriff, 1997, Hanks, 2000). Why, at the end, should we prefer to ascribe the occurrences of *anno luce* (light year) to the first rather than to the second sense? Is *anno luce* not composed of twelve months too? Is the temporal dimension of *anno luce* the most important to define it or is anno *luce* more a measure of time rather than a measure of space? Does it have anything to do with the fact that in IWN and in WordNet2.1 *anno luce* is defined as a measure of length and not of time? These *naïve* questions are just around the corner every time we want to exactly define the sense of a word.

But what we are trying to do here is not to define *the sense in abstract*, but rather to understand what the best sense organization is for the computational exploitation of the bulk of information stored in the resource. It is obvious that the answer to this question is not universally valid but it highly depends on the various final applications we are thinking of: the sense grouping required by Machine Translation will be inexorably different from the one required by our Question Aswering system (Kilgarriff, forthcomings). In the past years, a general tendency has emerged, i.e. considering the fine-grained sense distinctions proposed by computational lexicons (in particular by WordNet) too problematic for state of the art NLP. This tendency is clearly explained in (Ide and Wilks, forthcomings):

> *The question is [..] not whether NLP applications such as IR and MT need WSD (they do), but rather, what degree of disambiguation they need and whether or not pre-defined sense inventories can provide it.[…] NLP applications, when they need WSD, seem to need homograph-level disambiguation, involving those senses that psycholinguists see as represented separately in the mental lexicon, are lexicalized cross-linguistically, or are domain-dependent. Finer-grained distinctions are rarely needed, and when they are, more robust and different kinds of processing are required. […] for the purposes of NLP, work on the problem of WSD should focus on the broader distinctions that can be determined reliably from context.*

From this type of observation, a line of research originated, dedicated to the reorganization of senses grouping proposed by WordNet and WordNet-like lexicons. This effort is headed by the studies described in (Peters *et al*., 1998) and (Palmer *et al*, 2001), directed to the creation of coarse-grained clusters of WordNet senses.

The results of our work seem to confirm the general tendency that a more coarse-grained distinction among the senses of the lexicons is enough for the QA task. There are cases, however, that show how the distinction among even very close readings of the same word is somehow useful to the requirements of the application and the reason seems to be the fact that, as we said, QA can be defined as a "Named-Entity-sensitive" application.

The necessity of more underspecified sense distinction is obvious when we analyse the question *In quale anno Thomas Mann ha ricevuto il premio Nobel?* (In What year Thomas Mann won the Noble Prize?). In this case, we want our application to be able to derive the answer type YEAR>DATE in order to recognize the answer among the textual material returned by the Search Engine. The senses of *anno* in IWN all share

fundamental information: they are all hyponyms of {tempo, periodo} (time, period) and they are all subsumed by the same Top Concepts, i.e. TIME and QUANTITY. Only the second sense, defined as "*periodo di dodici mesi in genere*" (a generic period of twelve months) (the most general one), has few hyponyms and a meronym (*mese*, month). The rest of the senses are completely identical, only their glosses are different. This representation and organization of the distinction among the senses is of no use under the computational point of view: as a matter of fact, glosses (unless they are analysed and exploited to derive other explicit information) are just strings of text completely opaque for the automatic processor. Thus, even if in IWN there are four senses of the word *anno*, an automatic procedure will be unlikely to operate on them as separate senses. That is why the AT DATE has been connected to all the four synsets and a sort of super-sense of the word was created. This expedient also allows the system to overcome the case of incorrect automatic sense selection. The same strategy has also been applied to typical cases of *regular polysemy*, for example linking the node CITY of the AT Taxonomy to the two SIMPLE-CLIPS SemUs of *città* (city):

Città (*esteso centro abitato punto di riferimento del territorio circostante per amministrazione, economia, politica, cultura, ecc.*) (a large urban area…) -- IS-A: centro (centre) – Type: GeopoliticalLocation

Città (*la popolazione che abita in una citta'*) (the population of a city) – IS-A: popolazione (population) – Type: HumanGroup

In SIMPLE-CLIPS these cases are already connected by means of specific relation, the RegularPolysemy, but when exploiting the vertical links to derive the Answer Type it is surely simpler to connect both nodes to the ATTaxonomy. We think it is quite useful to collapse these cases of regular polysemy in *de facto* unique senses[77], since for our application this kind of polysemy does not seem to have any important impact on the analysis of the question and in the successive steps of retrieval and answer identification. The two questions:

In quale citta' la Mosella incontra il Reno? (In what town does the Mosel meet the Rhine?) (CLEF2004question#184)
Quale città è stata insignita della medaglia al valore civile? (What town does receive the medal for civic valor?)

are examples of the two readings of the word *città*: in the first one, the geographical dimension of meaning seems more important than in the second one, where the population of the city (and not its physical territory) is supposed to be the receiver of the medal. Nevertheless, there is no actual need to distinguish the two readings, since the strategy that should be triggered in the answer detection module is the same: looking in the candidate answer for entities of the type CITY>LOCATION satisfying certain conditions. For this reason we decided to make all the senses of *città* directly available in the ATTaxonomy, thus avoiding the possibility that an incorrect sense selection could prevent of understanding the expected answer type.

[77] Obviously, not all the cases of regular polysemy are similarly unnecessary for the application.

These examples would seem to suggest that a coarser granularity would be necessary and, as the case of a*nno* (year) shows, that senses that share the same Top Concepts and the same hyperonyms should be candidates to be treated as a single sense. The problem is that the situation is much more complex and the qualitative analysis of the results shows that in other cases the distinction between very close readings of the same word can be useful for the application.

These cases are highlighted by the contrastive evaluation of the performance of the system when exploiting the two lexicons, evaluation that allows the observation of the impact of the diverse vertical organization of the information and the different extension of the represented meaning. For example, in the case of CLEF2004question#188, *Di quale gruppo Teresa Salgueiro e' la cantante?* (Of what band is Teresa Salgueiro the vocalist?), differently from what happened for IWN, the system was not able to derive the AT HUMAN GROUP because the semantics of *gruppo* in SIMPLE-CLIPS is too generic to be captured by the portion of lexicon subsumed by the AT. As a matter of fact, while in IWN a specific synset was created just to gather the "social" groups, in SIMPLE-CLIPS no similar concept is available and the "social" groups are collected instead by a node of the Top Ontology. The system is thus instructed to exploit the Top Concept instead of the IS-A and, when the ATT is classified under the Type Human Group (like in the case of *associazione* (association), *squadra* (team) etc.), the AT is correctly derived. But when the ATT is simply *gruppo* (like in the case of CLEF2004question#188) it is not recognized as human group since the only SemU available (which covers both groups of people and of things) is directly linked to the Constitutive node (Fig. 61).

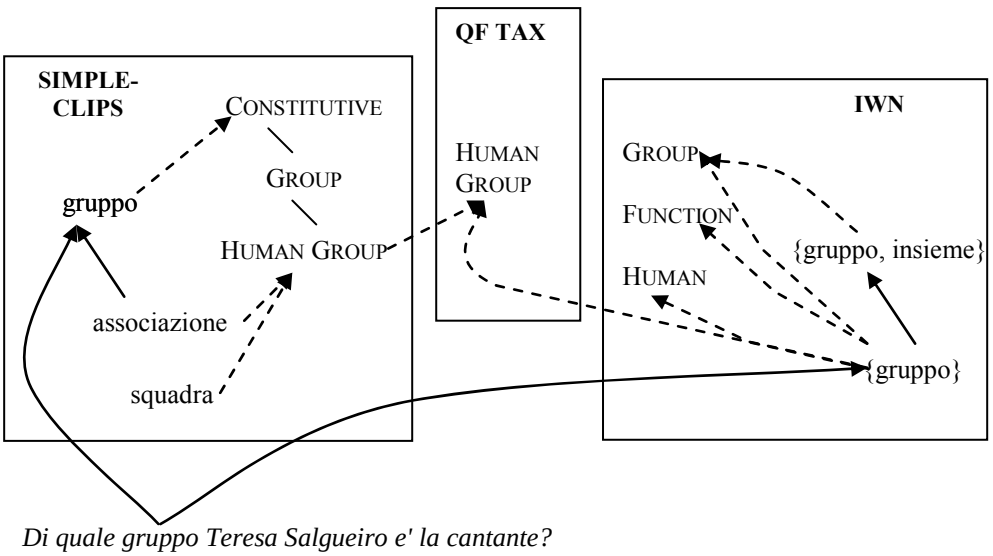


Fig. 61: derivation of the AT HUMAN GROUP in SIMPLE-CLIPS and in IWN

The case of the different encoding of *gruppo* in the two resources is however interesting. It begs the question of whether it was correct to isolate a sense of *gruppo* as composed only by people, distinguishing it by the more general sense of *gruppo* (that comprehends both people and abstract and concrete entities) that is also

encoded as its hyperonym. Another issue is then constituted by the attribution of the synonym *insieme* (set) to the more general sense of group and not to the "human" sense. The doubts about the legitimacy of such a sense distinction (that can also be found in printed dictionaries like the already mentioned Garzanti, 2005 and De Mauro, 2000) are well motivated: can we really state that there is a separate sense of *gruppo* that covers just the case of gathering of people or should we encode only the most general sense? Is it theoretically correct to represent such a specific sense like a hyponym of the more general one? From the point of view of our applicative exigencies we can say that such a sense distinction is surely worth being encoded: as a matter of fact, the more granular vertical organization allows the system to circumscribe a portion of the lexicon containing similar meanings (the various synsets *association*, *organization*, *team*, *political party*, *commercial enterprise*) and to infer from occurrences of *gruppo* similar to the one of CLEF2004question#188 that the system has to search for the answer among Named Entities of the type companies, teams etc.
gruppo 1 and *2* are similar for their ontological classification with *gruppo* 2: both share the same fundamental Top Concept (GROUP) but *gruppo 2* is more specifically described (it has more dimension of meaning).

gruppo 1 -- GROUP
gruppo 2 -- GROUP, FUNCTION, HUMAN

This example shows that even if a coarse-grained grouping of word meanings[78] is less problematic than the fine-grainedness present in our LRs, in some cases the distinction of two very close senses (even not theoretically well founded) can be appropriate for the exigencies of the application. In the case of QA, moreover, requirements concerning granularity is heavily connected to the distinctions that can actually be captured by a Named Entity Recognizer.

5.3 Breadth of the lexicon

The two lexicons provide the application with a reach repository of lexical senses. The vast majority of the words analysed by the system were in fact found in the lexicons, even with some exception due to the fact that SIMPLE-CLIPS is relatively smaller in size than IWN (cf. 1.1.3).
The two lexicons however differ for the support they provide in two specific cases, i.e. multiword recognition and exploitation of reflexive and transitive pronominal verbs.
We already said in 4.3.3.3.1 that about 16 question keywords should be considered not in isolation but rather as parts of multiword expressions (*bomba atomica*, atomic bomb, *campo di sterminio*, death camp, *salto con l'asta*, pole vault etc.). Most of these MWEs are listed among the lexical entries of the IWN database, while we can state that multiwords are not present in the current version of the SIMPLE-CLIPS lexicon, where

[78] The new grouping can be based on coarse ontological differences or even, as suggested by (Ide and Wilks, forthcomings), on fundamental difference as the ones between homographs.

only few, very general MWEs were introduced as dummy entries to help categorization of homogeneous sets of senses (*unità di misura*, unit of measurement, *essere umano*, human being, etc).

Recognition of poly-lexical units is an important sub-task, foreseen by most of the state-of-the-art QA systems. As far as our system is concerned, MWE recognition is important in the module for the assessment of keyword relevance, where *pena di morte* (death sentence) and *genere musicale* (musical genre) have surely a smaller number of hyponyms than the more generic terms *pena* (pain) and *genere* (genre). Recognizing MWEs is also of crucial importance in the module for AT identification (where analysing *unità di misura*, unit of measurement, is different from isolating the more general *unità*, unit) and during query expansion (where expanding *campo di sterminio*, death camp, with the synonyms of *sterminio*, i.e. *ecatombe*, *eccidio*, *macello*, *massacro*, *strage,* massacre, hecatomb etc., is not productive and creates noise while it would be useful to expand the multiword expression with *campo di concentramento,* concentration camp). Another issue that has to be taken into account is the possibility of actually exploiting MWEs that are encoded in IWN: as a matter of fact, in IWN MWEs are just strings of text with one or more blanks and no information is given on the internal structure of the entry. This prevents the system to easily morphologically analyse the various parts of the entry and in this sense handling the morphological variation of the keyword in the question and in the answer is not straightforward. For nominal synset this operation is less difficult since we can quite easily match occurrences by working on the endings of the single parts of the entry in order to manage the singular-plural alternation. On the contrary, working on the morphological variation of verbs is a much more difficult task and in IWN no information is given that may drive the analysis of verbal poly-lexical synsets.

If in SIMPLE-CLIPS basically no multiwords is present, also the way they are encoded in IWN cannot be considered optimal since no actual criterion has been adopted to decide what should be in the lexicon and what should not: many of the multiwords we find in IWN are semantically compositional and transparent, and are not the kind of frozen terms we think of when we talk about multiwords: *strumento musicale* (musical instrument), *area geografica* (geographic area), *bomba atomica* (atomic bomb) are all expressions whose meaning can be derived by the sum of the meaning of their parts. We however think that a higher acceptance of this type of expressions in the lexicon could have some very positive effects on the performance of the applications. For sure, however, the description of the syntactic structure of the entry is something that should not be missing from the synset: without a complete description of the internal structure of the mwe and without any clues about how it can vary in the target text no full exploitation of this type of information will be really feasible.

While IWN is a useful provider of multiword expressions, the SIMPLE-CLIPS lexicon is more suited to allow the system to analyse and exploit reflexive and transitive pronominal verbs. As a matter of fact, a substantial difference exists on the treatment of this type of verbs between the linguistic analysis chain (and the Treebank) and the IWN synsets: in IWN, the transitive pronominal and reflexive forms of the verb have been encoded in distinct synsets, as the example of *sposare-sposarsi* (to marry-to get married) shows:

{sposare, maritare, coniugare, congiungere -- *unire in matrimonio*} (to join in marriage)
{sposarsi, colvolare a nozze, coniugarsi – *unirsi in matrimonio con qualcuno*} (to get married to s.o.)

The output of the chunker, on the contrary, foresees the recognition of the basic form *sposare* and the encoding of the clitic. In SIMPLE-CLIPS, for which we do not have any problems of this type, a strategy coherent with the output of the *chunker* has been followed.
What makes not feasible the direct exploitation of the IWN entries of this type as such is the fact that in IWN no mechanism is foreseen to represent the "reflexivity" of the verb, and no representation is given of the internal organization of the string "sposarsi". A possible way to overcome this representational problem (that is also connected to the possibility of analysing and using the verbal multiword expressions) might consist in analysing all the verbs using the chunker, providing in this way a syntactically aware representation of the lexical entries.

One of the things that should be stressed is the fact that often what can be called a Named Entity is a multiword in its turn. Named Entities are not only the ones signalled by the NE recognizer but also some of the entries of type instance available in IWN and SIMPLE-CLIPS (respectively about 3500 and 1200). No clear criterion has been used to select the instance classes to insert in the lexicons, with the exception of availability in external repositories. As result, in both lexicons we find names of cities, countries and other instances. Nevertheless, fixed repositories containing only few thousands of instances cannot be considered an answer to the requirement of an application whose resulting answers are at more than 80% a named entity. Instances available in the lexicons were thus used only in very few cases (to confirm the answer detected by using syntax-based rules), all the times when the name of an important city was involved. All the other times, when the sought entities were a person's name, or the title of a movie, the name of a ship or others, the two lexicons did not provide any valid help to our application. Such repositories should not be internal parts of the lexicon but rather external repositories, like the Gazetteers usually exploited by QA systems (such as the *CIA World Factbook*[79], a database containing geographical, political, and economical profiles of all the countries in the world).

5.4 Depth of the Lexicon

The comparison of the two lexicons we made in Chapter 1 highlighted some important differences in the overall models and in the way information was acquired: differently from IWN, in SIMPLE-CLIPS we find a more clear adoption of the Generative Lexicon framework as theoretical model, the use of Templates as guidelines for lexicographers and, above all, the presence of predicates connected to sub-categorization frames. In IWN, on the other side, a methodology for multilingual linking has been defined, instantiated with equivalent relations to the Interlingual Index (as we said, in the SIMPLE-CLIPS model an alternative

[79] http://www.cia.gov/cia/publications/factbook/

multilingual setting is envisaged and described in details but not yet realized); but the most obvious difference is the strong stress, in IWN, on the notion of synonymy as *semantic glue* for concept definition.

However, even if some differences can be identified, the types of information that populate the content of the two lexicons are similar and comparable. In the next paragraphs, we will analyse the outcome of the exploitation of what we can call the depth of our lexicons, i.e. all the typologies of linguistic information expressed in their entries and by the connectivity among them.

5.4.1 Hyperonymy exploitation

By observing the whole application it is possible to see that the most exploited type of semantic information is *hyperonymy*. It is widely used in the module for the assessment of the relevance of the keyword, in the answer type determination and also in answer detection.

As regards the *specificity option*, two are the features that have been taken into account in the enhanced prototype: the specific/generic opposition and the belonging to taxonomies and types gathering meta-linguistic *word meanings*. IWN is able to provide a useful support in recognizing generic ATTs that should not be sent to the Search Engine. By using SIMPLE-CLIPS, on the contrary, it seems more difficult to exactly recognize generic ATTs since only two of the six generic terms were correctly identified.

Two of the generic ATTs that were not recognized by SIMPLE-CLIPS were also missed by IWN: *ingrediente* in CLEF2004question#52 (*Qual è un ingrediente base della cucina giapponese?*, What is a basic ingredient of Japanese cuisine?) and *scopo* in CLEF2004question #91 (Qual era lo scopo della prima azione sostenuta da Greenpeace?, What did the first action carried out by Greenpeace aim at?).

These word meanings are very interesting from our point of view: even if links driven by the *hyperonymy* relation are the most exploited in our prototype, they are however not completely reliable. In our experiment, we tried to establish an objective measure of the level of specificity, saying that, for a lexical item to be indicated as "vague" it has to gather a certain number of hyponyms and it has to be located in a relatively high position in the hierarchies. As usual, the notion of hyperonymy seems to work quite well with some parts of the lexicon, less with others: listing the hyponyms of particular animal specie is surely easier than listing all the possible *ingredients* or *aims*. Why? The first problem is that while there is always a finite set of living entities that can be categorized under a given animal or plant specie (all the *moths*, all the *dogs*, all the types of *fern* etc.), the items in the "ingrediente" and "scopo" sets are indefinite in their number. In the tale of *Snowhite*, the witch prepares her potion using the "smile of a rabbit" as an ingredient: should we list it among the hyponyms of *ingrediente*? Obviously not. The reason for this deep difference is that while in case of the animal and plant taxonomies the most important dimension of meaning is the classification on the base of some formal properties, in case of *ingrediente* and *scopo* the most salient dimensions are the telic and constitutive ones. When working on these types of word meaning, we should be aware that betting heavily on the information conveyed by the hyperonymy relation can not be the best idea. In what follow, we verify

whether the two concepts are represented in ItalWordNet and SIMPLE-CLIPS in a way that offers to the application the chance to work on them with appropriate strategies not based on hyperonymy.

The two semantic lexicons under analysis represent the word meaning *ingrediente* as an almost empty set: SIMPLE-CLIPS lists two hyponyms under the SemU *ingrediente*: the *bouillon cube* and *breadcrumbs*. No hyponyms at all under the equivalent *synset* in IWN. The problem is that in this way *ingrediente* is at the same level with *cibo* (food), *insetticida* (insect-powder), *cemento* (cement), *tintura* (dye) etc. and, as a co-hyponym of all these concepts, it is mutually exclusive with them: from this organization it logically derives that no food can be an ingredient. In this sense, the solution adopted in IWN is surely wrong and should be revised.

Differently from IWN and from SIMPLE-CLIPS, we can see that in WordNet2.1 about 250 *synsets* are classified as ingredients: going from *salt* to *basil*, from *Bolognese pasta sauce* to *anchovy paste* etc. In this case, it seems that under the ingredient node were gathered foods, garments, sauces etc. that usually are not directly consumed but rather mixed and composed in dishes. Even if this choice would be successful for the need to individuate generic word meaning in this specific case, this is surely not a good way to handle this type of inconsistency. As a matter of fact, looking at the actual answer to our question (*Qual è l'ingrediente base della cucina giapponese*? What is a basic ingredient of Japanese cuisine?), we learn that the basic ingredients are *pesce* (fish), *tofu* and *verdura* (vegetables):

Sabato si sottoporrà a nuovi controlli medici in un ospedale di Tokyo, ma sembra che il ritorno alla tradizionale cucina giapponese a base di tofu (pasta di fagioli), pesce e verdure gli abbia riportato salute e buon umore. (….the traditional Japanese cuisine based on tofu…fish and vegetables….)

As we will further discuss when we explore the results of the dynamic query technique, in the end what we would have needed was something that links (directly or indirectly) *tofu*, *pesce* and *verdura* to the ATT *ingrediente*. Exploiting the list of hyponyms available in WordNet would not help us anyway.

We thus see that also something very concrete, like food, can be difficult to categorize: *ingrediente* is a top node in the Constitutive Template in SIMPLE-CLIPS while in IWN it is represented as a food without any indication of its being a part (the same happens in WordNet2.1). The choice, made by IWN and by WN2.1, to encode *ingredient* as a co-hyponym of other substances or foodstuff (like for example *white rise*, *cocoa*, *flour* etc.) is however incorrect, since co-hyponyms, for the definition of hyperonymy itself, should be mutually exclusive while every hyponyms of substance can be also an ingredient.

Another inconsistency in the vertical lexical organization is observable looking at the way the various "ingredients" are classified in the two lexicons: while in SIMPLE-CLIPS the *bouillon cube* is an ingredient, in IWN it is represented as an extract (of beef) (thus as a product, thus as an object). The problem is that, as often happens in semantic lexicons, to describe all these different yet contemporary aspects of word meaning always subsumption, hierarchical relations are exploited (in the form both of canonical IS-A relation and inclusion in ontological nodes). Also applications often rely on vertical information to reach their aims. An

important change in the operative and representational practice would be starting to exploit the orthogonal dimensions of meaning as conveyed by semantic relations.

As we explained in Chapter 1, in SIMPLE-CLIPS the representational devices for overcoming the rigidity of the one-dimensional hierarchical structure of the lexicon are particularly advanced; in IWN, comparable devices are foreseen and in both resources lexical items are orthogonally ascribed to ontological concepts and described in terms of rich sets of semantic relations.

We can see that in the two lexicons the entry *dado* (bouillon cube) is described in quite a complex and rich way:

i) in SIMPLE-CLIPS, it is described as a part of a dish, focussing on its constitutive role (Pustejovsky, 1995);

ii) in IWN, it is described by focussing on the way it is created (by expressing its agentive role, represented by the ARTIFACT top concept and by the ISA relation targeting the *synset prodotto*, product) and on the basis of the intrinsic telic nature of the top concept FUNCTION (Pustejovsky, 1995).

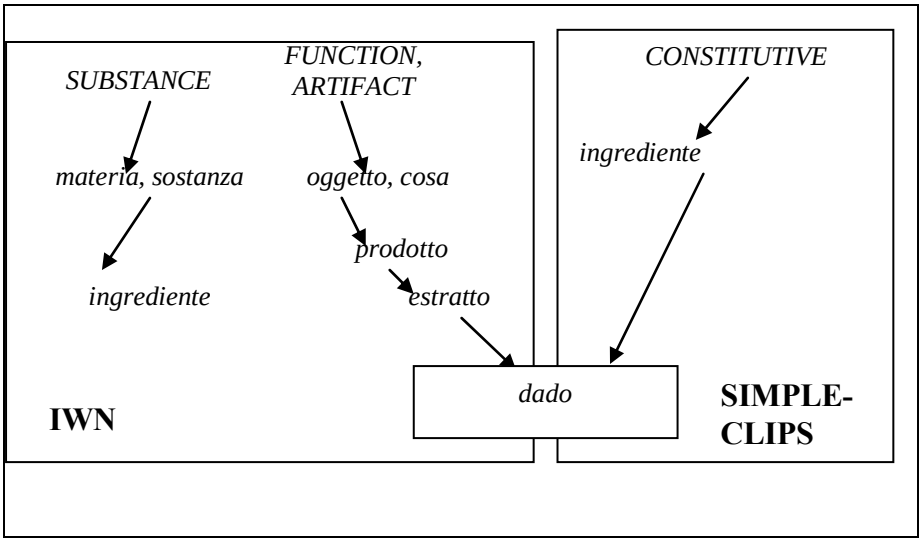


Fig. 62: the semantic content of the lexical entry *bouillon cube* in SIMPLE-CLIPS and IWN

In this way, the constitutive, agentive and telic nature of the lexical entry is however always conveyed by vertical and hierarchical links and not by "horizontal" relations. These semantic representations do not seem enough to allow an automatic procedure to identify the "generic" nature of the concept *ingrediente*, since the adopted strategy uses the number of hyponyms and the depth of the corresponding taxonomy as its only measures.

What seems to happen is that the IS-A relation has become a sort of repository of different aspects of meaning, aspects that collapse into the same label losing their important distinctions. Important reference for

this kind of considerations is the work done by the Guarino and Gangemi's research group and resulted in the OntoClean methodology (Gangemi *et al.*, 2001a, Gangemi *et al*., 2001b).

The OntoClean methodology is the characterization of ontological categories in terms of formal meta-properties based on the fundamental distinction between *individuals* and *concepts*. Formal properties in the OntoClean approach are *rigidity*, *identity*, *dependence*, *types* and *roles*, *extensionality*, *concreteness*, *unity*, *singularity*, and *plurality* (the interested reader can find a detailed description of each property in Gangemi *et al*., 2001a). One of the problems raised by analysing WordNet with OntoClean is what is called the *ISA overloading* phenomenon (Gangemi *et al*., 2001b and Guarino, 1998)[80]: ISA is often intended as a lexical relation between words, which not always reflects an ontological relation between classes of entities of the world. This generates problems such as:

i) confusion of senses (a window is both an artefact and a place),

ii) reduction of sense (a physical object is an amount of matter, an association is a group),

iii) overgeneralization (a place is a physical object)

iv) type-to-role link (an apple is both fruit and food).

What mostly effects the problem emerged in the analysis of *ingrediente* is the confusion between *type* and *role* that can be recognized in IWN.

In OntoClean, *Type* and *Role* are formal meta-categories defined by means of multiple meta-properties: a *Type* is a rigid property (i.e. essential to all its instances) that supplies an identity criterion (i.e. not inherited by any subsuming property) and is not notionally dependent on another property. A *Role* is instead an anti-rigid property that is notionally dependent. It is a material role if it carries (but not supplies) an identity criterion and a formal role otherwise. In this sense, *person* would be a type, *student* a material role and *part* is an example of formal role, since it carries no identity and is notionally dependent.

The linguistic design of the two semantic lexicons is surely open to represent transversal dimensions of meaning (the telic, agentive and constitutive roles) mainly by means of semantic relations and ontological classification. Nevertheless, in our computational lexicons (n particular in IWN) there is an over exploitation of the ISA expressive means, used to express *purpose*, *function*, *origin*, *material*, *part-whole information* etc.

Probably, the most coherent and logically valid solution would be to find a representational device capable of stating that "all substances can be ingredients if they are used to prepare dishes or medicines" and to precisely recognize the telic and constituency dimension of *ingrediente*. The representation of the concept *ingrediente* closer to this type of solution is the one proposed in SIMPLE-CLIPS, where *ingrediente* is a SemU without hyperonym, directly connected to the Top Ontology node Constitutive (Fig. 63).

[80] Difficulties connected to the semantics of the ISA relation were already emerged during the ACQUILEX Project (Calzolari, 1991; Calzolari et al., 1993).

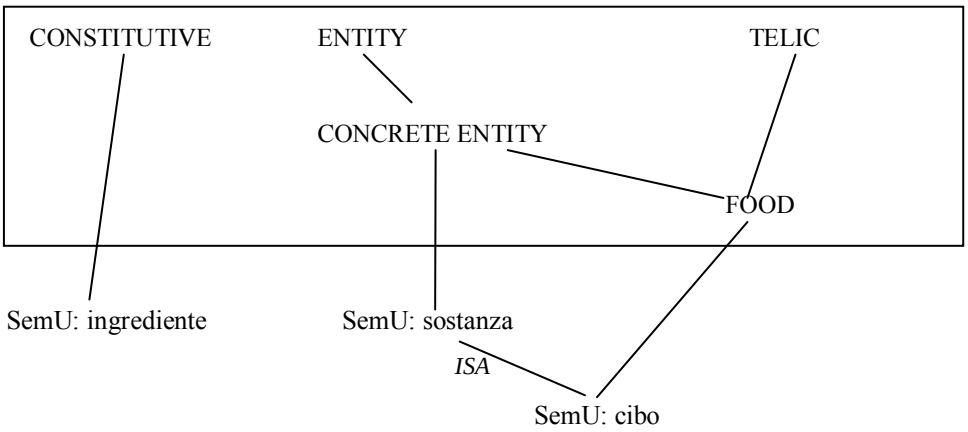


Fig. 63: Representation of *ingrediente* (ingredient), *sostanza* (substance) and *cibo* (food) in SIMPLE-CLIPS.

Unfortunately, in the current version of SIMPLE-CLIPS nothing links the concept *ingrediente* to the taxonomy of substances. However, the linguistic model provides a semantic relation that can be established to represent that link, the *Used_As*, a relation of the telic role. If such a connection would be available, a possible strategy may consist in exploiting the inheritance mechanism of the IS-A relation and in making the concept *ingredient* become a sort of attribute of all the substances in the semantic net by means of the specific telic semantic relation. Using Gangemi *et al*.'s words: we should distinguish between the *type* and the *role*, preserving the ISA vertical structure for the types (the actual types of substance, like *dust*, *cement*, *grease*, *food* etc.) and allowing for an orthogonal account of their telic dimension (Fig. 64).

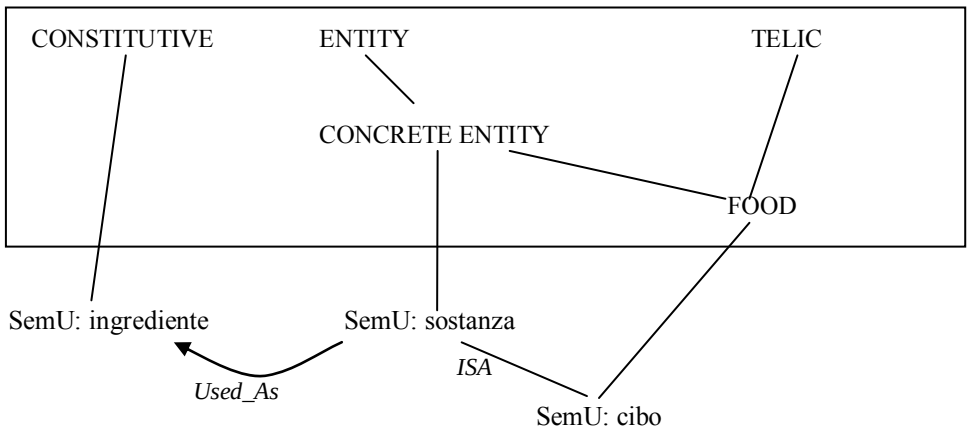


Fig. 64: Establishing a link between *ingrediente* (ingredient) and *sostanza* (substance) to support inference

A representation as the one showed above would be the prerequisite for new strategies to be implemented in the application. The general idea would be, in case of ATT strongly connoted by a telic or constitutive dimension, not exploiting the ISA relation but rather a lexical chain (in this case the one formed

by the *Used_As* and by the *ISA* relation) to have an idea of which and how many are the concepts in the lexicon interested by the "property" *ingrediente*. Together with its being a root of the taxonomies in SIMPLE-CLIPS, this representation would give an idea of the vagueness of the concept *ingrediente* as well the possibility of retrieving the possible ingredients in the other modules of the application, like the dynamic query formulation. What it would be useful is the possibility to have the SemU *ingrediente* classified under the TELIC and not only the CONSTITUTIVE semantic type. In this way, the system may follow a strategy consisting in reading the Semantic Type of the ATT and, in case of Semantic Type TELIC, not applying the strategy based on the ISA exploitation but rather on the lexical chains supported by semantic relations of type telic (Fig. 65). As a matter of fact, not every type of relations would be useful in this strategy but only those expressing the concept "this property can be applied to this set of concepts".

Obviously, the same strategy can be adopted also in ItalWordNet, by making *ingrediente* become a very high concept in the hierarchies, directly linked to the FUNCTION and PART Top Concepts of the ItalWordNet Top Ontology and adding to the model a semantic relation allowing the representation of telic information not only between a second and a first order entity (for these cases the ROLE/INVOLVED_INSTRUMENT relation is available) but also between two first order entities (like *ingrediente* and *sostanza* are). Moreover, the ItalWordNet linguistic model allows also overcoming a problem that emerges with SIMPLE-CLIPS. As a matter of fact, while IWN allows the contemporary attribution of a word meaning to two distinct Top Concepts (in this case, the PART and FUNCTION Top Concepts), in SIMPLE-CLIPS the encoder has to choose which, among the highest in the Ontology, is the node more appropriate to represent the fundamental dimension of meaning of the lexical entry. To represent *ingrediente*, for example, the encoder has to choose between TELIC and CONSTITUTIVE and is not allowed to select both.

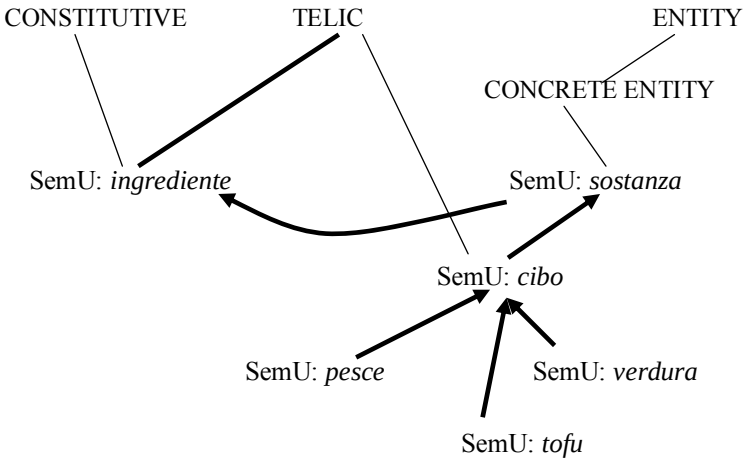


Fig. 65: lexical chains tracing a useful inferential path in SIMPLE-CLIPS.

The explicit constitutive dimension of *ingrediente* may be used in a different way, for example by exploiting the *meronymy* relations in case of questions like *Quale ingrediente è usato nella preparazione della ragù?* (What is the ingredient of the ragù sauce?). In that case, the system can look for a meronymy link between *ragù* and *carne* (meat) or *pomodoro* (tomato) directly in the lexicon.

The solutions discussed above would be obviously aimed at further enforcing the formal and logic solidity of the representations expressed in language resources. What is sure, however, is that the logic consistency of the lexicon does not constitute the ultimate solution to the exigencies of language processing. As a matter of fact, all the considerations made for *ingredient* are not valid for *scopo* (aim): everything can be identified as an aim, any action, any concrete object, and any condition. It is not possible to establish a useful semantic relation between *scopo* and other concepts in the lexicon. It is easier to think to a dedicated, ad-hoc strategy that, for example, read the ATT *scopo*, look in the paragraphs for constructions of the type *fare X per Y* (to do X for Y). In this sense, it seems that an approach based on corpus exploitation would be more fruitful.

There are other cases in which SIMPLE-CLIPS missed the identification of a generic ATT: *professione* (profession) and *categoria* (category). Both are examples of alternative classification of identical lexical items in different language resources: all the entries classified as *professione* in IWN, in SIMPLE-CLIPS are categorized as activities (*notariato,* profession of notary), discipline (*giornalismo,* jounalism) etc. Nevertheless, we can see that the exploitation of the IWN hyponyms of *professione* in the dynamic query module was not successful since, among the various hyponym, we do not find the names of professions (*giornalista*, journalist, *agente*, agent, *panettiere*, baker etc.) that are instead classified as *lavoratore* (worker, employed). We will see that the way the profession taxonomies are organized in the lexicons has some negative effect in the way they are exploited in the module for the creation of dynamic queries.

What is classified in IWN as *categoria* is organized under the SemU *gruppo* in SIMPLE-CLIPS. In both cases of *categoria* and *professione*, we can see that the tendency of SIMPLE-CLIPS to propose flatter taxonomies is confirmed, with many lexical items attached directly to the highest nodes in the hierarchies.

SIMPLE-CLIPS failed also in handling the specific ATT *quotidiano*. While this kind of publication is the leaf of a 8-level deep taxonomy dedicated to textual material (*quotidiano* > *giornale* > *edizione*, *pubblicazione* >…> *oggetto* > *entità*), in SIMPLE-CLIPS it is only at the 3rd level of a taxonomy that has *insieme* (*set*, *group*) as root.

Unfortunately, the tendency in SIMPLE-CLIPS to concentrate the hyponyms under few, generic nodes has a double negative effect: taxonomies are too flat and SemUs often have no hyponyms in a way that thwarts the possibility of exploiting measures like the indication of the level of vagueness.

In 4.3.3.4 we described another method for assessing the salience of the ATT which is to determine whether it belongs to meta-linguistic taxonomies or templates (the taxonomy headed by {unità_linguistica} in IWN and the Template METALANGUAGE in SIMPLE-CLIPS). This is the case of the various *nome* (name), *titolo* (title), *cognome* (surname), etc. Both resources were a valid support in the identification of

such ATTs, with only one exception: the ATT *titolo* that is encoded not as belonging to the METALANGUAGE template but to INFORMATION.

Another module where hyperonymy is heavily exploited is answer detection, where the lexicon helps the system to recognize clues that fulfil the informative needs expressed in the question. The tasks are two: i) the substitution of the Answer Type Term with its hyponyms in the composition of the query, ii) the expansion of the lexical occurrences on which the various syntactically-based rules apply (as illustrated in 4.3.5). In the next paragraphs we show how also these modules are impacted by problems in the vertical organization of the concepts.

In section 4.3.3.4.6 we listed almost twenty questions introduced by interrogative adjective *Quale* and *Che* (What, Which…) by the interrogative pronoun *Quale* and by other "ambiguous" interrogative forms of the type *dammi* (give me..), *dimmi* (tell me), *nomina* (name…) etc. for which the system did not find any Answer Type or any Answer Type correspondent to a named entity class. When using ItalWordNet, the exploitation of the all-level hyponyms of the answer type term is often effective, and the system is able to generate the query with the candidate answer that leads to the extraction of the answer paragraph. This is true, for example, for questions like Qual *è l'unità di misura di frequenza?* (What is the frequency unit?), *Come vengono chiamati i piloti suicidi giapponesi?* (What are Japanese suicide pilots called?), *Che lingua si parla in Germania?* (What language is spoken in Germany?) etc.

Nevertheless, sometimes, as it happens when the system has to derive the Answer Type for *ingrediente* and *scopo*, the exploitation of the ISA relation shows its points of weakness. There is a fundamental problem concerning what is hyperonymy and when and where we should encode it. The already (when we analysed the ATTs *ingrediente* and *scopo*) discussed problem of *ISA-overloading* is at the end a matter of ambiguity: different conceptualizations are confused by using a same semantic relation. The ambiguity, however, invests not only the semantic lexicon under analysis but also the adopted searching strategy itself. As a matter of fact, sometimes we find in the lexicon not logically consistent links like the one between *ingrediente* and *sostanza* of IWN as well the ones between *ingrediente* and *pangrattato* of SIMPLE-CLIPS. But often what has to be revised is the idea itself that the link between *ingrediente* and *tofu* in the sentence:

il tofu è un ingrediente della cucina giapponese

can be decoded as a hyperonymy. In that case, what is weak is the practice of always decoding as hyperonymy the relation between the ATT and the answer.

If we have a look at many ATTs of questions for which the dynamically generated queries did not help to pin-point the answer, we can see that we arrive to the same conclusions we got when we analysed the case of *ingrediente* and *scopo* and their exploitation in the modules for Answer Type identification and assessment of keyword relevance, i.e. that it is the notion of ISA itself that cannot be successfully adopted in these cases:

q_9: *Quale incarico ricopre Ariel Sharon?* (What office does Ariel Sharon hold?)

q_31: *Qual è la professione di James Bond? (*What is James Bond's job?)
q_52: *Qual è un ingrediente base della cucina giapponese?* (What is a basic ingredient of Japanese cuisine?)
q_55*: Di quale nazionalità erano le petroliere che hanno causato la catastrofe ecologica vicino a Trinidad e Tobago nel 1979?* (What nationality were the two oil tankers that caused the ecological catastrophe near Trinidad and Tobago in 1979?)
q_91*: Qual era lo scopo della prima azione sostenuta da Greenpeace?* (What did the first action carried out by Greenpeace aim at?)
q_94: *Qual è un fattore di rischio per le malattie cardiovascolari?* (What is a risk factor for cardiovascular diseases?)
q_95*: Quale è la categoria professionale più a rischio di cancro ai polmoni?* (What professional category is more at risk of lung cancer?)
q_98: *Dammi un sintomo con cui si presenta l'affezione da virus Ebola.* (Give a symptom of the Ebola virus.)

Between the ATTs of the "failed" questions and the corresponding answers what we really want to find in our lexicons is not the ISA relation but rather other types of relations, more "in line" with the fundamental dimensions of meaning of the Answer Type Term. As a matter of fact, it is correct that there is not a hyperonymy relation between the following ATTs and the actual answer:

1. *Incarico -- ministro degli esteri* (office – Foreign Minister)
2. *Professione -- agente segreto* (profession – secret agent)
3. *Ingrediente -- tofu* (ingredient – tofu)
4. *sintomo– Diarrea* (symptom– diarrhoea)
5. *fattore di rischio– Fumo* (risk factor –smoking)
6. *nazionalità – liberiana* (nationality – Liberian)

etc.

In the same way, *impedire* (*impedire gli esperimenti nucleari americani sull'isola di Amchitka, nelle Aleutine*) is not a hyponym of *scopo*, *mal di gola* and *febbre* are not inherently symptoms but rather pathological conditions (like domain-specific ontologies[81] suggest), *ipertensione* and *fumo* are not *fattori di rischio* per se. In this sense, it is the strategy adopted by the system that is not designed in a granular enough way: when dealing with these complex types of word meanings, what should be exploited is not the formal dimension but rather the telic, constitutive and agentive dimensions. At this point, in order to trigger alternative strategies as the one suggested for *ingrediente,* the system has to find in the lexical entry some explicit elements signalling the the case should not be treated by exploiting hyperonymy (in that case, we suggested that the useful element was the Telic Semantic Type). If we look at all the "failed" ATTs (Table 23), we see that in SIMPLE-CLIPS and IWN none of them is defined by recurring to the formal role (with the exception of *ingrediente* in IWN):

[81] MESH: www.nlm.nih.gov/mesh/2005/MeSHtree.html, and UMLS: www.nlm.nih.gov/research/umls/umlsmain.html.

ATT	SIMPLE-CLIPS Semantic Type	IWN Top Concept
Incarico (office, duty)	purpose Act, feature: Telic	Social, Agentive
Professione (profession, job)	purpose Act (unification Path-relational Act), feature: Telic	Static Unbounded Event Social Agentive Purpose
Ingrediente (ingredient)	constitutive	Substance
Sintomo (symptom)	phenomenon, supertype Event	phenomenal dynamic
fattore (factor)	agentive	function
Nazionalità (nationality)	property	static property

Table 23: ATTs and their SIMPLE-CLIPS Semantic Types and IWN Top Concepts

After having recognized the "special status" of these ATTs, the challenge would be exploiting available semantic relations to automatically support the reasoning that is needed to pinpoint the answer. If we look at the exemplification of semantic paths provided in Chapter 3, we see that in those cases some available "ways" to go from the ATTs and the answer can be identified.

Nevertheless, the exploitation of the available information is everything but simple and as the result of this research we can state that probably finding a general, systematic strategy to handle these questions is not feasible at the current state of the art. First of all, we see that also at ontological level the two resources give different interpretation of the word meaning: what is an Agentive entity for ItalWordNet (*incarico*, office) is a Purpose Act in SIMPLE-CLIPS. This complicates the implementation of systematic strategies aimed at handling the semantics of these lexical entries. However, we can try to develop strategies based on a single language resource. When considering ItalWordNet, we see that *incarico* is classified as an Agentive situation, i.e. a situation in which "..a controlling agent causes a dynamic change; e.g. to kill, to do; to act" (Rodriguez *et al*., 1998). A possible strategy may consist in exploiting the agent/role type of relations to retrieve the "agent" the question is looking for. As we can see in Fig. 66, this type of connection is not directly available in the lexical entry of the ATT *incarico* but it has to be "calculated" by taking into consideration the inheritance mechanism triggered by the hyperonymy connecting *incarico* to *lavoro* (work) and *impiegato* (employed) and *ministro* (minister).

Fig. 66: the semantic path connecting *incarico* (office) and *ministro* (minister) in ItalWordNet

This is not a insurmountable problem, even if the practical implementation of a similar strategy is not simple. The real problem is that a similar strategy would not allow a systematic treatment of all the "Agentive" ATTs: in case of *Qual è un fattore di rischio per le malattie cardiovascolari?* (What is a risk factor for cardiovascular diseases?), looking for the relation of type "involved/role_agent" is not useful at all. We are again in the same situation determinate by the analysis of the ATT *scopo*: as a matter of fact, everything can be a factor of risk for something, as well as an aim can.

A particularly difficult situation is represented by the exploitation of the subset of professions in IWN and SIMPLE-CLIPS in the case of questions of the type "*quale è la professione/l'incarico/l'ufficio…?*. In the test bed we find the two questions *Quale incarico ricopre Ariel Sharon?* and *Qual è la professione di James Bond?*. The answers are respectively *ministro degli esteri* and *agente segreto*, but we can see that neither in IWN nor in SIMPLE-CLIPS we can find them listed among the hyponyms of *professione-incarico:* in IWN the list of professions is organized (as it happens in WordNet) solely as a taxonomy having as root the synset {*persona*, *essere umano*, *uomo*, *individuo*}, with an intermediate level represented by the synset {*lavoratore*}[82]. Following the OntoClean recommendations (Gangemi et al, 2001a), we should avoid to encode the various professions under the Human node, since a role (the profession) cannot be subsumed by a type (the human being). If this recommendation would have been followed, the system would have been able to exploit the subset of lexicon dedicated to professions to individuate the answer to this type of questions. Nevertheless, we are not persuaded that a simple shift of the taxonomy from the human to the activity node would have been completely resolutive of the problem. As a matter of fact, even if the professions organized under the node professione>attività can be exploited in the answer detection phase, the

[82] {lavoratore} is somehow a fictitious sense because, even if lavoratore (worker) is fully an Italian concept and word, none would call a spy a worker. The presence of this intermediate node of the hierarchy grouping such an heterogeneous set of concept is only determined by the necessity to isolate the hyponyms of human being characterised by their performing an activity for which they are paid for.

classification under the HUMAN node is useful when we want to derive the Answer Type. Consider the following questions:

CLEF2004question#87: *Quale presidente americano è stato renitente alla leva?* (What American president failed to report for military service?)
TRECquestion#1338: Who is the actress known for her role in the movie "Gypsy"? (*Quale attrice è conosciuta per il suo ruolo nel film "gypsy"?*)
TRECquestion#967: What American composer wrote the music for "West Side Story"? (*Quale compositore americano compose la musica di "west side story"?*)

In these cases, it is of primary importance to allow the system to *understand* that the answer is probably a person's name (and in this sense both lexicons meet the exigencies of the application). Differently, an Answer Type ACTIVITY would have not been of any help. In our perspective, the casting out nines constituted by usefulness in application is obviously of great importance to verify the correctness of a choice thus, in this specific case, the fact that two distinct yet specular modules of the same system require two different classification of the ATT is very problematic.

Also under a more theoretical point of view it seems that the interpretation of professions as "types of human" has certain validity. For example, it is possible to see how it passes the diagnostic test of hyperonymy (like the following, used in the EuroWordNet project):

General test:

yes	a	A/an X is a/an Y with certain properties It is a X and therefore also a Y If it is a X then it must be a Y
no	b	the converse of any of the (a) sentences.
Conditions:		- both X and Y are singular nouns or plural nouns

Test applied to the exemplificative link teacher-person:

	a	A teacher is a person with certain properties
	b	?A person is a teacher with certain properties
	a	It is a teacher and therefore also a person
	b	?It is a person and therefore also a teacher
	a	If it is a teacher then it must be a person
	b	?If it is a person then it must be a teacher
Effect:		teacher N HAS_HYPERONYM person N person N HAS_HYPONYM teacher N

It is moreover interesting to note that the test fails if the tested hyperonym is *activity*:

	a	A teacher is an activity with certain properties
	b	?A activity is a teacher with certain properties
	a	It is a teacher and therefore also an activity
	b	?It is a activity and therefore also a teacher
	a	If it is a teacher then it must be a activity
	b	?If it is an activity then it must be a teacher
Effect:		teacher N HAS_HYPERONYM person N person N HAS_HYPONYM teacher N

Trying to decide which hyperonym is better, we should take into consideration a third possibility, i.e. that both *activity* and *human* may be valid hyperonyms. As a matter of fact, in our opinion, even if maybe never discussed in the literature on the subject (Apresjan, 1973, Pustejovsky, 1995), the human-profession alternation can be studied as a particular case of *regular polysemy*. The polysemy between the activity itself and the person who performs it emerges when we analyse the two occurrences of *insegnante*:

Gianni è (*un*) *insegnante* (Gianni is a teacher)
Gianni fa l insegnante (*Gianni does the teacher)

The difference between the two senses is also signalled by the different article used in the two sentences. In IWN, this polysemy is not expressed, while an echo of it can be found in the way professions are organized in SIMPLE-CLIPS, i.e. by recurring to a sort of "transversal" and "hybrid" Semantic Type[83], Profession that is however subsumed by the more general Type Human. We talked about a hybrid Type because the label Profession mimics the lexical concept *profession*, an abstract concept denoting an activity, but still it inherits from its SuperType Human its "ontological truth", the feature of concreteness.
At the end, we can see that the possible strategies are:

- encoding professions as role (as recommended by (Gangemi e al., 2001))
- encoding professions as human (like in IWN, WN and SIMPLE-CLIPS)
- encoding professions as a case of regular polysemy, thus foreseeing two taxonomies

In our opinion, none of these possibilities can be considered an ultimate solution: the first one has as consequence that the Answer Type Human cannot be identified; the second one, that, as it happens in our system, the hyponyms cannot be generated starting from the ATT *professione*; the third one implies that the system has to differently consider the sense of the noun denoting profession when it is in the question (*Quale insegnate ha vinto il premio…?*) and when it is the answer (*l'insegnante Mario Rossi*…). Moreover, a proliferation of senses is something that we surely do not want in our lexicon.

In such a difficult situation concerning what senses should be encoded and with what hyperonym, an automatic system has however to find its way out: in the specific case of questions asking about *professions*, *jobs*, *offices* etc., what can be exploited is that in IWN the ROLE/INVOLVED_AGENT relations connect in many points the two taxonomies of i) humans performing a job, and ii) jobs and activities. We repropose the figure (Fig. 67) that shows the path connecting *professione* and *agente segreto*.

[83] With some exceptions and internal inconsistencies: *agente* (agent) is classified as Profession, while *spia* (spy) under Human.

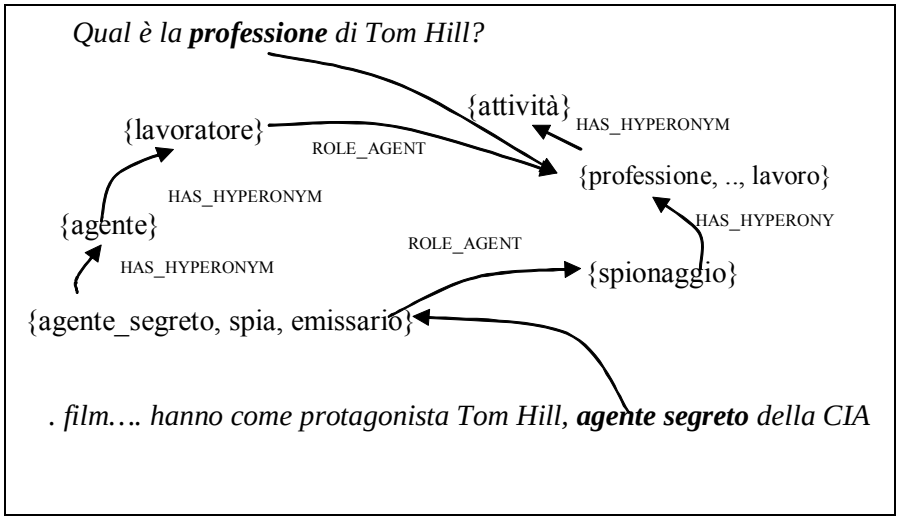


Fig. 67: IWN relations connecting *professione* and *agente segreto*

The most practicable solution is implementing a more sophisticated approach able to generate the hyponyms of the synset {lavoratore} when the ATT is *professione.* Notwithstanding questions asking about profession are very frequent, the challenge would be being able to individuate not an *ad-hoc* strategy for handling them but rather a general schema to resolve all the cases for which the ISA relation is not a valid way to individuate the answer.

The situation, as we already showed in 3.2.1, is different in SIMPLE-CLIPS, where we cannot find any common points between the SemU *agente* and the SemU *professione*. Again, only an ad-hoc strategy seems at the moment practicable, since the system only in the case of ATT *professione* (and *incarico*, etc.) has to exploit not the hyponyms of the ATT itself but all the SemUs classified under the Semantic Type *Profession*.

5.4.2 Difficulties in concretely implementing lexical chains.

In the previous paragraphs, we often mentioned the need to expand the scope of the analysis from the immediate semantic context of the lexical item at hand to a broader set of word meanings indirectly linked to the lexical item by means of more complex chains. For example, we showed that in case of *ingrediente*, instead of exploiting the normal hyperonymy relation, a concatenation of the *used_as + hyperonymy* relation seemed more useful to achieve the sought results. In the same way, when we wanted to generate the list of professions, what was needed was a chain constituted by the *involved/role_agent + hyperonymy* relation.

We already widely discussed about the difficulties (that seem insurmountable) to find solutions of general significance when the fundamental dimension in the description of the word meaning is not the formal one. Nevertheless, we also showed that, by isolating some cases, some preferred paths can maybe be tempted. But, even if these paths could be determined for some cases, an even more complex problem would emerge: as a matter of fact, Open-Domain Question Answering is an application that presupposes a fundamental step,

i.e. the reduction of the problem complexity by exploiting the Search Engine, which is used to return only a subset of all the paragraphs of the document collection. This step is really important, because the ambition of this type of QA is just the possibility to work on a virtually open-ended collection of documents, as the Web is. Thus, every time we indicate a possible semantic path among, for example, the Answer Type Term and the answer, we operate an adulteration of the situation the application will have to handle in reality: a paragraph containing words completely different from the terms of the query will not be taken into account by the application if an expanded query did not return it[84].

Everything we said about these lexical chains connecting question and answer should be thus implemented as an expansion of the query, where the original queries are enriched and augmented with terms encountered while navigating through the semantic paths. In case of chains constituted by a fixed number of semantic relations this can be quite easily implemented in a system. This is for example what we did when we expand the query using the concatenation of *PolysemyNationality-LivesIn* relations when using SIMPLE-CLIPS to derive pair of adjective-name of country (Albanese-Albania, italiano-Italia) (cf. 4.3.4). Differently, if the concatenation is not made of a fixed number and type of relations, the system will difficultly handle the expansion procedure: for example, if we look at the case of question *Quale è l'incarico di Ariel Sharon?* (What is the office of Ariel Sharon?), we see that it is not easy for the system to iteratively compose the query that, progressively, incorporates new terms and without really knowing in advance when and where to stop this iteration. Fig. 68 shows the list of lexical entries that should be iteratively added to the query in order to handle this question.

Fig. 68: terms that should be involved in the query expansion to derive the answer to questionCLEF#9.

[84] Luckily, most of the time, the question and the "answering" paragraph share at least one term, an this allows the system to work on the semantic content of the returned paragraphs in an efficaciously way. Sometime, the threshold of 40 paragraphs we chose as maximum number of paragraphs to be analysed by the system determines that some useful paragraphs can however be discarded.

In those cases, the only final control that can be hypothesized is the "found" statement, i.e. the fact that the answer is finally found. It goes without saying that, as we already said in 4.3.4., a similar strategy is not of easy implementation, because there is the case that the Search Engine returns documents that contain the "intermediate" terms without being the answer to the question. Luckily, even if the system was not able to completely exploit this information, it was however able to answer this question by recurring to syntax-based rules[85].

5.4.3 Disjoint conceptual representation in SIMPLE-CLIPS.

In SIMPLE-CLIPS, the possibility of exploiting the content of the SemU is complicated by the disjointed representation of the concepts, that are fragmented in different SemUs instead of being unitary represented in a groups of synonyms. The exploitation of the hyponyms should then take into account this specificity, in some way collecting not only the hyponyms of the specific variant corresponding to the ATT but also all the hyponyms of the synonyms of the ATT. In this way it would be possible to generate the name of the German monetary unit (*marco*) as a candidate answer to CLEF2004question#24, *che moneta si usa in Germania?* starting from the ATT *moneta* (Fig. 69).

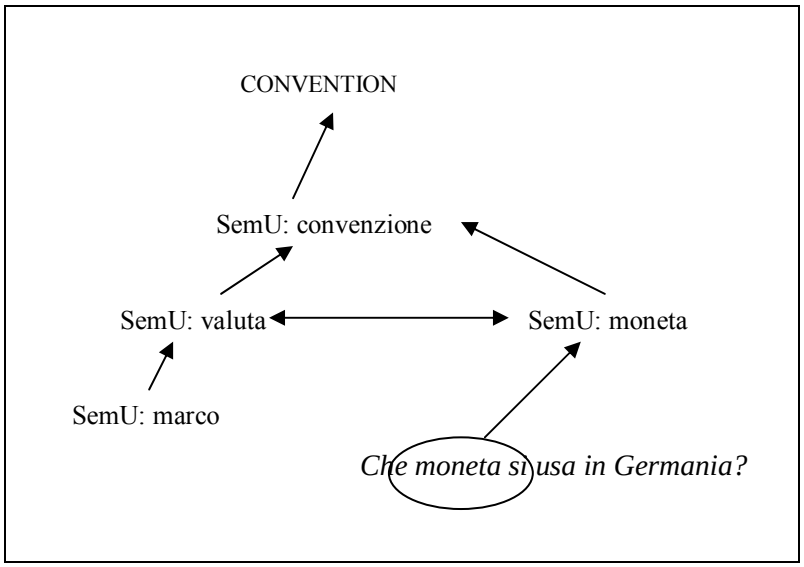


Fig. 69: the taxonomies describing monetary values in SIMPLE-CLIPS

[85] For example the one based on the fact that often the office-profession of someone is expressed as an adposition of the subjet.

5.4.4 Lexical semantic expansion of syntax-based rules

In chapter 4 we described the strategy of lexically expanding the cases to which the syntactic-based rules apply. There are some case where the strategy is actually efficacious: in the case of IWN, for example, it helped the system to pinpoint the answer to the question used as example in Chapter 4 and in the questionnaire of Chapter 3, i.e.:

CLEF2004question#7: *Quanti membri della scorta sono morti nell'attentato al giudice Falcone?* (How many escorts were killed in the assassination of Judge Falcone?)

In that case, the system exploited the hyperonym of the word *membro* (member), i.e. the synset {persona, individuo, uomo, essere umano} (person, individual, human being) to match the question with the candidate answers:

….dove furono uccisi il giudice Giovanni Falcone, la moglie Francesca Morvillo e ***tre uomini*** *della scorta*
…nella quale morirono il giudice Giovanni Falcone, la moglie Francesca Morvillo e ***tre uomini*** *della scorta.*
….strage di Capaci in cui morirono il giudice falcone, la moglie ***e tre uomini*** *della scorta*
….quella strage di Capaci che costo' la vita al giudice Giovanni Falcone, alla moglie Francesca Morvillo, a Rocco Di Cillo, Antonio Montinari e Vito Schifani, ***tre uomini*** *della scorta.*

However, other relevant answers for the same question were not identified: this is true, for example, for candidate answers:

…furono uccisi il giudice Giovanni Falcone con la moglie Francesca Morvillo e tre agenti della scorta.
…con vittime Falcone, la moglie e ***tre poliziotti*** *che li scortavano,*
….della strage di Capaci, dove morirono il giudice Giovanni Falcone, la sua compagna Francesca Morvillo e ***tre degli agenti*** *di scorta.*

In these cases, the lexical variants of the word in the questions are *agente* (officer) and *poliziotto* (policeman), that in IWN belong to the same synset and are (2nd level) hyponyms of {persona, uomo, essere umano, individuo}. These synsets are represented in a way that make them logically co-hyponyms of *membro* (member), thus logically mutually exclusive with it. Again, a logical inconsistence can be detected in this type of representation, since obviously a policeman can also be member.

This question were instead failed by SIMPLE-CLIPS, where the SemU *membro* and *uomo* are represented in a completely unrelated way (see Fig. 70). In the same figure is moreover possible to see how a

second sense of *uomo* is however present, the one referring to "male human being", that cannot be anyway exploited since it is co-hyponym of *membro*.

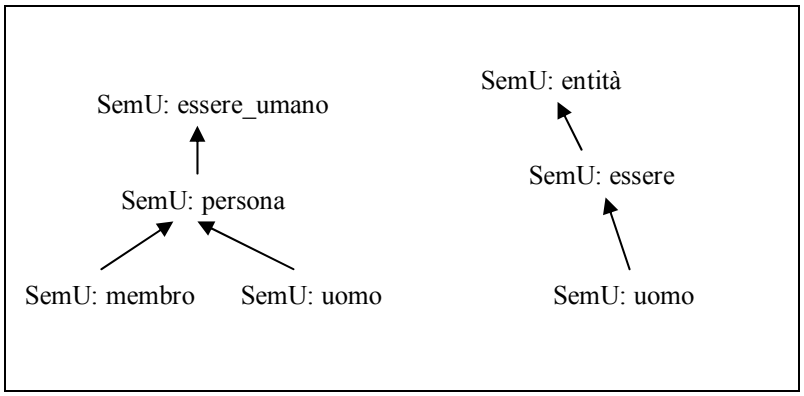


Fig. 70: representation of *membro* (member) and uomo (human being) in SIMPLE-CLIPS

Another example of question answered by the system by exploiting IWN but not SIMPLE-CLIPS is CLEF2004question#94: *Qual è un fattore di rischio per le malattie cardiovascolari?* (What is a risk factor for cardiovascular diseases?)
The answer were in the paragraph:

Il dato e' preoccupante, soprattutto in considerazione del fatto che l'ipertensione rappresenta un importante fattore di rischio per le malattie cardiovascolari. (….hypertension represents an important risk factor for cardiovascular diseases…)

The mismatch between question and answer is in the verb, that in the question is *essere* (to be) and in the answer is *rappresentare* (to represent). But while in IWN a sense of *essere* belongs to the same synset of *rappresentare*, in SIMPLE-CLIPS no path connects the two verbs.

Another failure is represented by the case of CLEF2004question#130: *Che cosa ha influenzato l'"effetto Tequila"?* (What did the "Tequila Effect" influence?). The answer says:

L'"effetto tequila" messicano, infatti, ha gia' seriamente danneggiato molti mercati latinoamericani, mettendo in fuga capitali che sono invece fondamentali per la crescita economica dei paesi della regione. (The "Tequila Effect" has already damaged many Latin-American market…)

In this case, in order to match question and answer, a link should be established between the verb *influenzare* (to influence) and *danneggiare* (damage) (if something damages something it means that influenced it), but

this connection is not supported by IWN and SIMPLE-CLIPS. The models of both lexicons, however, would have allowed the encoding of a relation linking the two word meaning. At the same time, however, it is legitimate to ask ourselves why a lexicographer should have encoded such a link, that is not something that seems prototypical of the meaning of the two entries.

5.4.5 Parts and Wholes

We already discussed about some difficulties emerging from the exploitation of the taxonomies of "*gruppo*" (group). The case of *group* is also interesting because one of the points of weakness of the exploitation of LRs is just *mereotopology*, i.e., as reported in (Gangemi et al., 2001a), the connection of two core theoretical tools for formal ontological design: the theory of parts and of wholes.

If we look at the taxonomy of the most general sense of *gruppo* (gruppo 1) in IWN we can see that it is an amalgam of very heterogeneous concepts (more than 1400). Among the first level hyponyms we find words like:

{*imbracatura*} – sling
{*sciame*} -- swarm
{*convoglio*} -- train
{*bendaggio, fasciatura*} -- bandage
{*contabilità*} -- bookkeeping
{*attrezzatura, equipaggiamento, dotazione*} -- equipping
{*squadra, squadriglia*} -- squad
{*paniere*} – basket of goods
Etc.

The same happens in SIMPLE-CLIPS, where we can see that different types of lexemes are classified as hyponyms of the SemU *insieme* under the Semantic Type GROUP:

Minutaglia-- bits and pieces
Scuderia -- stable
Terminologia -- terminology
Sporco -- dirt
Rettile -- reptile
Tubazione -- tube
Scogliera -- rocks
Simbolismo – symbolism
Segnaletica – system of sign
Etc.

Many lexemes are in these taxonomies only "thanks" to the content and form of their lexicographic definition. As a matter of fact, if we look at the definition of *imbracatura*, we can see that it is defined as "the set of ropes used to sling", *bendaggio* is "a set of bandage", *contabilità* is "the set of books and accounts of an organization", *rettile* is a "class of animal" etc. But in this way we lose the fundamental dimension of meaning that constitutes the backbone of such concepts: all these word meanings are not simply sets or

groups, but rather they are physical objects (the *sling* and the *bandage*), activities (the *bookkeeping*), animal (the *reptile*). Again, we are probably in front of what (Gangemi et al., 2001a) describes as a case of *IS-A overloading*, i.e. the phenomenon of *reduction of sense* according to which "the ISA link points to an aspect of the meaning of a given concept that does not fully account for its identity". In this case, the attribution of the ISA is clearly due to the practice of identifying in the genus term of the definition the hyperonym of the *definiendum*, regardless of the loss of information.

Moreover, even when the "constitutive" dimension of meaning is very marked, it can be the case that the hyperonymy link to {insieme, gruppo} is not what is needed to derive the Answer Type capable of matching question and answer. This is the case, for example, of CLEF2004question#137: *Dammi il nome di una catena di Fast Food* (Name a fast food chain). The system was not able to derive the Answer Type because in IWN the ATT *catena* is represented as a group[86]. The synset {catena} is also linked by means of a HAS_MERO_MEMBER relation to the synset {società, impresa, azienda, ditta, compagnia} but the problem is that the answer itself is something that in the text is identified as a *company* (and not as a *group*):

"Solo qualche centinaio di dollari invece per un certificato-regalo della catena di fast food Mc Donald's, che Elvis regalo' per scherzo al cugino Billy Smith, reca un valore nominale di soli 50 cen" (….of the chain of fast foods McDonald's…)

Fig. 71 provides a graphical description of the way the concept *catena* in represented in ItalWordNet. It is easy to see how the horizontal account that IWN does of the "company" meaning of chain is not useful within the adopted strategy of AT identification based on the exploitation of the subsumption relations in LRs with respect to an ontology of types.

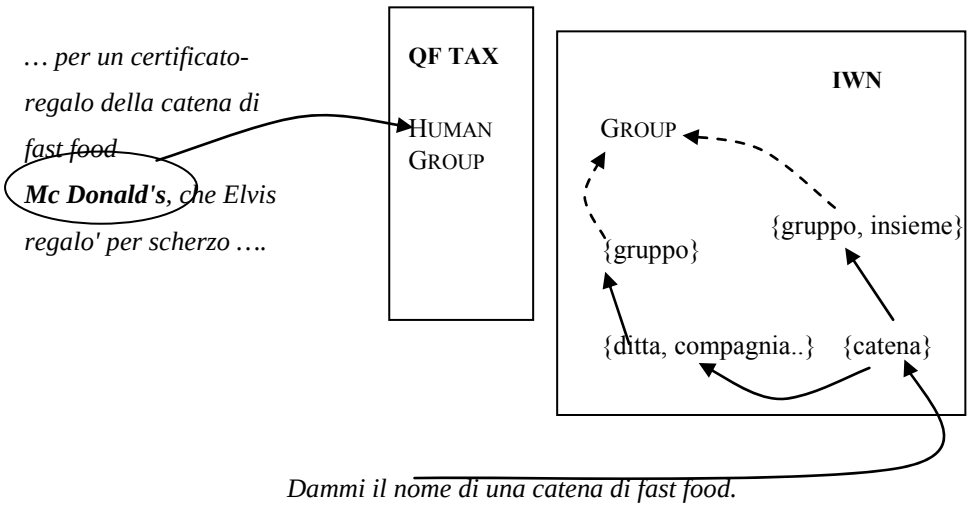


Fig. 71: missing the "company" meaning of chain in IWN

[86] In SIMPLE-CLIPS the commercial sense of *catena* (chain) is missing.

A different representation, consisting of ascribing *catena* to the synset {azienda, ditta, compagnia} (company) would have easily allowed the recognition of the AT HUMAN GROUP. In that case, however, the two alternative hyperonyms should be seen as quite different ways of interpreting the *merotopology* of *catena*: when the hyperonym is {insieme, gruppo}, what the representation conveys is that *catena* is a group of companies (or stores). If the hyperonyms were {gruppo 2}, the constitutive dimension would be due to the fact that the companies are composed by groups of people.

The exploitation of the taxonomies dedicated to the representation of parts is not clear either. In the CLEF2004 test bed there is only one question with ATT *parte* (part):

CLEF2004question#96: *Dammi il nome di una parte dell'organismo attaccata dal virus Ebola* (Name a part of the body that is affected by the Ebola virus)

A specific Answer Type, Body Parts, was foreseen in the AT Taxonomy. But it has been connected to the Body_Part Type of SIMPLE-CLIPS and to the synset {parte_del_corpo, parte_anatomica} (body part, anatomic part) of IWN. The problem is that in the question we find the ATT *parte dell'organismo* (part of the organism) and in this way we lose the possibility of exploiting the established link. Thus, not matching the ATT of the question with the "right" senses in the lexicons, the system is neither able to derive the correct AT nor to exploit the hyponyms of the ATT in the queries dynamically formulated. As far as IWN is concerned, this situation is due to an inconsistency in the encoding of the entry: as a matter of fact, in the lexicon the synset {organismo, corpo} (organism, body) is present. When having to encode the meronyms of *corpo*, however, the node {parte del corpo, parte anatomica} was created to vertically organize all the meronyms. But the word *organismo*, encoded as synonym of *corpo*, disappeared from the new synset, thus not allowing the correct recognition of the ATT of the question (Fig. 72).

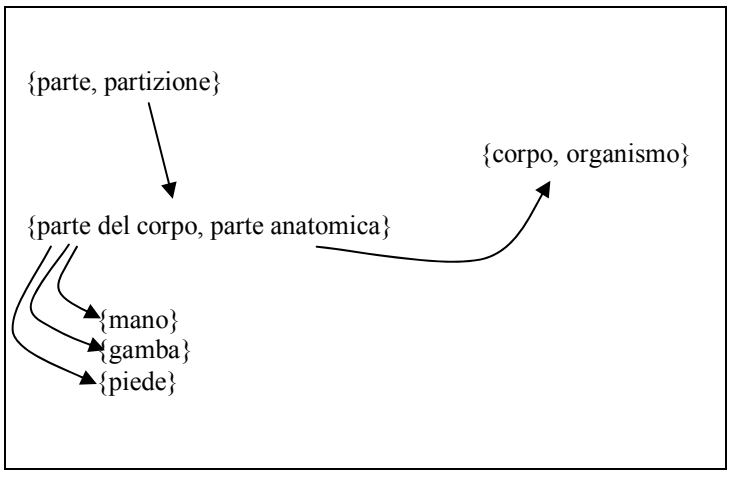


Fig. 72: connectivity of the synset {parte del corpo} in ItalWordNet

What is quite interesting, however, is that a HAS_MERONYM relation is encoded between {corpo, organismo} and {parte del corpo, parte anatomica}, thus it has been possible to adopt the specific rule-based strategy to match the entry of the semantic net and the ATT described in 4.3.3.4.6.

5.4.6 Loops in hyperonym chain

One of the things that most invalidate the exploitation of taxonomical information is the presence of loops. Unfortunately, in two cases:

CLEF2004question#36: *Che scuola frequenterà William, il figlio maggiore del principe Carlo?*

CLEF2004question#42: *In quale genere musicale si distingue Michael Jackson?*

the derivation of the AT by using SIMPLE-CLIPS was not possible due to loops in the taxonomical chain of the ATT *scuola* and *genere*.

5.4.7 Type Taxonomy

As far as question#42 (CLEF2004question#42: *In quale genere musicale si distingue Michael Jackson?*, In what music genre does Michael Jackson excel?) is concerned, IWN does not provide a valid support in AT identification. As a matter of fact, as we already illustrated in 4.3.3.4.5, no link has been established between the ATTaxonomy and the taxonomical portion with root *genere*, *tipo*, *sorta* (genre, type, sort) etc. because of its vagueness. A strategy to derive the "real" ATT is thus been studied (4.3.3.4.5) and, thanks to this, the ATT *genere musicale* was successfully exploited in the "dynamic query" module. Nevertheless, the necessity of a strategy consisting of exploiting the "type" word when it is present with its modifier as a lexical entry of the LR is a symptom of the problems and inconsistencies of the taxonomies that have as roots the synset {genere, tipo, sorta, fatta,specie, qualità} and the SemU *genere*.

The encoding of hyponyms of these lexical entries is mainly due to an incorrect interpretation of lexicographic definitions and it is particularly present in IWN because in this lexicon a more extensive use of semi-automatic extraction of information encoded in printed dictionaries has been applied. For example the definition of *pop* was:

pop: *genere musicale nato alla fine degli anni 60* (pop: music genre born at the end of the sixties)

the hyperonym *genere* (or the multiword *genere musicale*, as appears in IWN) was chosen to represent the synset. But when a different form of definition was preferred by the lexicographer, as, for example, in case of:

jazz: musica di origine afro-americana (jazz: music with Afro-American origin)

the chosen hyperonym is different. Thus, very often, similar entries too are treated and classified in a very different way. In the case of our question (CLEF2004question#42: *In quale genere musicale si distingue Michael Jackson?*), the answer was *pop*, and for this reason it was successfully identified by exploiting IWN, but what if the expected answer had been *lirica*, or *rock-and-rol*, that are classified directly under the hyperonym *musica*?

Even if the inconsistencies in hyperonymy attribution and taxonomic organization are surely elements of weakness of the lexicon, we have to admit that a rich connectivity can help to overcome problems deriving from such inconsistencies. This means that, even if the various types of music (pop, jazz, classic etc.) are organized under the two hyperonyms *genere musicale* and *musica*, this inconsistency can be overcome by making the two hyperonyms be connected by a semantic relation or by making them variants of the same synset, recognizing them as synonyms. This would help the system to answer in the same way both the question of the test-bed (In quale genere musicale si distingue Michael Jackson?) and the hypothetic question with ATT *musica*: *Quale musica suona Michael Jackson?* (What music does Michael Jackson play?). In ItalWordNet, no relation is established between *genere musicale* and *musica* and this does not allow the system to exploit such a kind of connection.

5.4.8 Decoding the expected answer type in questions introduced by Quanto

In the fourth chapter we discussed about the necessity to apply systematic strategies to precisely identify the Answer Type in the case of questions introduced by the interrogative adverb *Quanto*[87] (How much..). The questions, grouped according to the answer type we think they should have, are:

Answer Type: Height

M ITA 0175 Quanto è alto il monte Everest? (How tall is the Mount Everest?)
F IT IT 0190 Quanto è alto il K2?(How tall is the K2?)
M ITA 0190 Quanto è alto il Matterhorn? (How tall is the Matterhorn?)
<TREC-10>Quanto è alto il Sear Building? (How tall is the Sears Building?)
<TREC-10>Quanto è alto il Gateway Arch a S. Luis, MO? (How tall is the Gateway Arch, S. Luis, MO?)

Answer Type: Time

F 0046 IT IT Quanto ci vuole per andare da Londra a Parigi attraverso il tunnel della Manica?(How does it take to go from London to Paris through the tunnel in the English Channel?)
<TREC-10>Quanto vive in media un grillo? (What is the life expectancy for crickets?)
<TREC-10>Quanto dormì Rip Van Winkle? (How long did Rip Van Winkle sleep?)
<TREC-10>Quanto dura la gestazione di un elefante? (For how long is an elephant pregnant?)

[87] In the CLEF2004 test bed there are few cases of this type, thus we decided to analyse the total test bed constituted by the questions used in three CLEF editions and by the translation of the TREC-10 collection.

Answer Type: Weight

F 0095 IT IT Quanto pesa un quark top? (How much does a quark top weigh?)
<TREC-10>Quanto pesa il cervello di una donna adulta? (How much does the brain of an adult woman weigh?)
<TREC-10>Quanto pesa l'acqua? (How much does water weigh?)

Answer Type: Money

F 0138 IT IT Quanto costa il telefonino più economico sul mercato? (How much does the most expensive cell phone cost?)
F 0155 IT IT Quanto frutta allo stato italiano ogni anno la vendita di sigarette? (How much does the Italian State yearly earn from selling tobaccos?)
<TREC-10>Quanto costava un biglietto per il Titanic? (How much did a ticket for the Titanic cost?)
F 0126 IT IT A quanto ammontano le perdite subite dalla Barings? (How much did Barings lose?)

Answer Type: Length

F 0178 IT IT Quanto è lungo il confine tra Cina e Mongolia? (how long is the border between China and Mongolia?)
<TREC-10>Quanto dista Denver da Aspen? (How far is it from Denver to Aspen?)
<TREC-10>Quanto è lungo un miglio nautico? (How far is a nautical mile?)

Answer Type: Quantity[88]

F IT IT 0005 Di quanto aumenta la popolazione mondiale ogni anno? (How much does the world population increase each year?)
F IT IT 0012 A quanto ammonta il numero dei profughi palestinesi che si sono rifugiati in Libano? (How many are the Palestinian refugees in Lebanon?)
F IT IT 0074 A quanto ammonta la popolazione degli USA? (What does the USA population amount to?)
M ITA 0066 A quanto ammonta la popolazione mondiale? (What does the mondial population amount to?)
<TREC-10>quanto spesso l'Old Faithful erutta al parco nazionale di Yellowstone? (How often does Old Faithful erupt at Yellowstone National Park?)

The analysis of the contribution of LRs is discouraging: the system is able to determine in a precise way only the AT of questions for which a pattern matching on the syntactic form was foreseen, while for all the other questions the derivation of the expected answer type is everything but simple. In the case of questions:

Quanto frutta allo stato italiano ogni anno la vendita di sigarette? (How much yield Italy the cigarette commerce?)
A quanto ammontano le perdite subite dalla Barings? (How much did Barings lose?)

[88] In this last group we listed heterogeneous examples that could be further classified.

it is not in the verbs modified by the adverb that the system can find the distinctive feature able to discriminate between a generic AT Quantity and a more meaningful AT Money, but rather in the subject of the question, i.e. *vendita* (selling) and *perdita* (loss).

In ItalWordNet, even introducing a specific rule to analyse the subject of the question, the system would not find anything in the content of the two lexical entries able to trigger the Money Answer Type. We thought there was the possibility of exploiting the belonging of the lexical entry to POSSESSION, a Top Concept defined in (Rodrieguez *et al*., 1998) as collecting situations involving possession: static situation (*have*, *possess*, *possession*, *contain*, *consist of*, *own*) but also dynamic changes in possession, like *sell*, *buy*, *give*, *donate*, *steal*, *take*, *receive*, *send*. But, differently from these concepts, the only Top Concept assigned to the synset *perdita* (loss) is PART, while *vendita* (differently from its XPOS_NEAR_SYNONYM *vendere*) is categorize solely under the Top Concepts DYNAMIC and CAUSE, without any other component of meaning.

In SIMPLE-CLIPS the situation is slightly different: the SemU for *perdita* is currently not encoded in the lexicon but the SemU *vendita* is correctly categorized under a specific Semantic Type, Transaction, that can be used to select the Money Answer Type.

Particularly difficult are the questions for which the system should derive the AT Time:

Quanto ci vuole per andare da Londra a Parigi attraverso il tunnel della Manica?
Quanto vive in media un grillo?
Quanto dormì Rip Van Winkle?

The first question is actually ambiguous because the expected answer may be also of type Money. The next two questions require the decoding of the temporal dimension of the meaning of the verbs *vivere* and *dormire*, that is completely missing from their semantics as it appear in the two lexicons.

In general, however, what is most discouraging is not the impossibility of analysing these types of question, but rather the fact that even more "simple" questions, like the ones with the form [*alto*|*pesante*|*lungo*|*costoso*|*distante*] and *Quanto* + [*pesare*|*durare*|*costare*|*durare*|*dista*] cannot be correctly analysed without recurring to ad-hoc lexical-based rules. As a matter of fact, no meaning components can be systematically and reliably exploited as a fixed feature that drives the interpretation of the question.

In chapter 4 we expressed the wish to find a method to derive these Answer Types in a systematic way without recurring to the ad-hoc rules encoded in the baseline prototype, but this was not the case.

5.4.9 Exploitation of Xpos relations

The evaluation of the adoption of query expansion methods in our prototype is not simple. If a small improvement can in fact be detected, we have to recognize that the benefits deriving from the use of synonyms and other related terms is really modest. One of the things that minimize the impact of the use of LRs is the adoption of the stemming technique. As a matter of fact, most of the time, the information conveyed by the xpos semantic relations in IWN and by the predicate object in SIMPLE-CLIPS is not really

useful because the stemmer has already correctly identified and extracted the root of the word, thus enabling the retrieval of occurrences not only morphologically but also semantically correlated. Thus, there was no need to send to the Search Engine the two keywords consegnare and consegna (expressed in IWN and SIMPLE-CLIPS respectively by means of a xpos_synonym relation and a predicate) since the system already uses the stemmed keyword *consegn**, that includes both. This does not happen always: in the case of *invasione* and *invadere*, for example, the stemmer provides two different stem (*invas** and *invad**) and in this case the support from the resources was not useless.

Stemming and exploitation of semantic information in this way are concurrent strategies to obtain the same results; the problem is that using stemming techniques is much simpler and more straightforward than navigating through the SemUs and synsets of our resources, collecting correlated items and disambiguating word senses. Obviously, the Search Engine has to support the search with truncation of the keyword but this is the case of most of the Search Engines available today.

In general, this can be said only for links between different parts of speech: in IWN, the XPOS_SYNONYM and AGENT/PATIENT_ROLE/INVOLVED relations, in SIMPLE-CLIPS the *EventVerb, DeverbalNounVerb, StateVerb, ProcessVerb, AgentVerb*, *PatientVerb* relations and the SemUs connected to predicate via a Master, VerbPastParciple, AgentNominalization, PatientNominalization, ProcessNominalization typeOfLink.

There are, however, important differences between the two resources, in particular regarding the way in which the information we decided to exploit is actually encoded. All the *xpos* relations listed above and also the others connecting adjective and noun are encode with much more consistency in SIMPLE-CLIPS that IWN. This is not true, however, for synonymy that, even if foreseen in SIMPLE-CLIPS, was exploited only on 4 of the 55 questions. As can be imagined, synonyms were better exploited when using IWN, a lexicon that is based on the notion of synonymy itself.

5.4.10 Synonymy

Query expansion is more useful in the exploitation of *synonyms*. In the case of CLEF2004question#71 (*Quanti stati in America hanno la pena di morte?*) we can see that, in order to retrieve the paragraph with answer:

Albany (New York), 7 mar (ats/ansa/reuter) Lo stato americano di New York si è aggiunto oggi agli altri 37 stati USA che hanno ripristinato la pena capitale dopo che entrambi i rami del parlamento federale hanno approvato il provvedimento.

the submission of the synonymic multiword expression *pena capitale* instead of *pena di morte* was useful to retrieve the answer. The same happened for CLEF2004question#44 (*Chi è l'inventore del televisore?*), where the answer can be found only by submitting the synonym *televisione*.

Nevertheless, even if in IWN synonymy is much more encoded, we observe the presence of synonyms that are unlikely to be found in the corpus as lexical variant of the word in the question: in IWN, for example, we find the synset {donna, femmina} to express the meaning of *essere umano di sesso femminile*. The definition of *femmina* in the dictionary (Garzanti, 2005) (i.e. *essere umano di sesso femminile; donna, bambina*) seems to confirm this choice, but it is highly improbable that these two words are found and both used in the same context. In the case of CLEF2004question#143 (*In quale anno, prima del 1995, si è tenuta la Conferenza mondiale delle donne*) we do not want to expand the term *donna* with *femmina*, because the only context in which we think the two words would appear would be texts with a negative exception of the concept, like the one by Petrarca reported in (Garzanti, 2005) (*Femina è cosa mobil per natura*). The validity of the synonymy expressed in IWN is thus quite uncertain, also because it is encoded without any restrictions (for example by means of a NEAR_SYNONYM and a code "pejorative" on the variant).
In the same way, looking at all the questions where the word *mondo* (world) appears,:

Qual è la compagnia cinematografica più vecchia del mondo? (Qual è la compagnia cinematografica più vecchia del mondo?)
Chi è il più grande gestore di telefonia mobile al mondo? (Who is the biggest mobile phone operator in the world?)
Quante sono le religioni monoteiste nel mondo? (How many monotheistic religions are there in the world?)
Quanti sono i cattolici nel mondo? (How many Catholics are there in the world?)
Quanti sono gli ebrei nel mondo? (How many Jews are there in the world?)

we can see that it is very improbable to find it substituted by the synonyms provided by IWN (i.e. *globlo terracqueo, globo terrestre,* terraqueous globe, earth's globe). The exploitation of synonymy seems somehow a double-edged weapon: very often, in fact, the synonyms indicated in the synset determine more noise than the effective retrieval of new pertinent paragraphs. For some very frequent forms it seems that an approach consisting in ad-hoc, dedicated rules (for example stopping the expansion when the system find the word *mondo*) would be more advantageous.

5.5 Final Remarks

Before trying to draw some conclusions, we recapitulate the aims of our research as they have been individuated and discussed in the introduction.

The first, most "superficial" goal was to understand whether and to what extent the information encoded in computational lexicons can be exploited to support exigent information management functions, like the ones required by Open-Domain Question Answering. The first aim was thus to establish whether computational lexicons determine an improvement in the performance of the systems that use them.

The second, more subtle motivation of this dissertation was instead to enter in the specific tasks of the application in order to analyze from close-up what the practical exigencies of the system are and whether the information available in computational lexicons can efficaciously enter in its mechanism. Open-Domain Question Answering was just chosen among other applications because it is a complex task, constituted by a number of more specific modules concerning "semantics-aware" retrieval and analysis of information. The modules of the application consist of: i) lexico-semantic analysis of the question and the answer texts, ii) retrieval phase and iii) individuation of the paragraphs semantically closer to the question. QA is thus an application that involves a series of quite basic steps of analysis that can concur to many final applications and in this sense it is particularly interesting. Entering in each of these steps allowed us to observe how language resources play their role and what their points of strength and weakness are with respect to the task at hand. The conclusions of this dissertation should be not only the mere ascertainment of the presence or not of an improvement in the performance of the application but also a final evaluation of what information cannot be exploited and why.

In the Introduction we raised an issue that we consider important, i.e. what does it mean to evaluate a semantic lexicon? The reason of such a question is somehow connected to the "ambiguity" of the term *lexicon* itself. As a matter of fact, a lexicon is a multi-faceted object consisting of a set of lexical entries, a model and a representational framework. In the light of the analysis of the obtained results, we can now try to list the cases of lexicon deficiency that seem to be at the base of many system failures:

a) the lexicon does not provide a specific information where it should (problem of lexicon coverage);
b) the lexicon provides an evident "wrong" information, derived by an error in the encoding practice;
c) the specific linguistic model does not allow the representation of the "useful" information but it could be changed and improved in order to do it;
d) the lexicon provides the required information but in the system only very granular, almost ad-hoc strategies to handle the specific cases can be developed;
e) the system does not find in the lexicon the support it needs and it is not possible to figure out any symbolic representation able to overcome this limit.

These situations present different level of complexity. Situations b) is the easiest to handle. The presence of evident "mistakes" in the knowledge base is something that cannot be completely eliminated but that can be corrected and limited with a good encoding practice, by paying particular attention to consistency and uniformity during the construction of the lexicon. Even very serious problems, such as loops in the hyperonym chains (like the ones reported in 5.4.6) can be corrected and it should be stressed the importance of dedicating part of the work during lexicon development to the definition of strategies to semi-automatically detect errors and inconsistencies.

Also situation c) can be quite easily overcome, since the exact identification of something that is missing in the model can be the first step towards an improvement and enrichment of the model itself (this is valid at least for the cases we met, for instance for the possibility to express, in SIMPLE-CLIPS, multiple

fundamental semantic dimensions of very high concepts by letting them have more than a single role, cf. 5.4.1).

Problems of coverage (situation a) cannot be dismissed that easily: even if the human encoder puts all the attention in encoding information for a lexical entry, this attention will never adequately provide the "entire" range of information that might be useful for working on the lexicon and operate inference. The analysis of the semantic paths connecting question and answer carried out during the examination of the results of the questionnaire (Chapter 3) shows that the connectivity available in the two lexicons is often not enough to support the exigencies of the system; sometimes the application suggests that a given relation between two word meanings would have supported the inference required by the task at hand, nevertheless that link does not seem so important and "prototypical" to be encoded in the lexicon (this happens often for links representing entailment and cause). Nevertheless, we know that the augmentation of large-scale lexicons is a very costly and time-consuming task. In this sense, important efforts are made to simplify and make possible such enrichment by resorting to techniques for the extraction of information from corpora (Hearst, 1992, Riloff and Shepherd, 1997, Berland and Charniak, 1999, Mann, 2002, Poesio et al., 2002). This is surely the most relevant direction to enrich computational lexicons. Also the notion of *interoperability* seems nowadays important to overcome limits and costs of lexical enrichment; for example, in (Soria et al., 2006), a concrete framework for the semi-automatic integration and interoperability of lexical resources is described.

The most difficult to evaluate are situations d) and e). In the first case, we said that even if some semantic paths can be identified to drive the required inference, the strategies the system can implement are too specific (like the almost ad-hoc rules for handling *ingrediente* –ingredient- and *professione* –profession- introduced as an alternative to hyperonymy exploitation in 5.4.1).

Example of situations in which it is difficult to even figure out a representation that would suit the specific task (situation e) are instead the ones requiring the treatment of answer type terms like *scopo* (aim) and *effetto* (effect) (discussed always in 5.4.1) or the recognition of Answer Types in case, for example, of questions introduced by ambiguous stems like *Quanto* (How much..) (see 5.4.8). For a certain classes of phenomena, like the possibility to identify "aims" in a collection of documents to answer questions like *Quale era lo scopo della prima azione sostenuta da Green Peace?* (What was the aim of the first action of Green Peace?), we think that a strategy based on corpus extraction would be really more efficacious than one based on language resources exploitation.

By taking into consideration all these different aspects and levels of complexity, some general conclusions can be drawn. First of all, we can state that a significant improvement in the performance of the system can be clearly recognized (cf. 5.1): lexico-semantics information does play a positive role on the results of the system. This result is in line with those described in (Harabagiu *et al*., 2001).

Nevertheless, only a small part of the knowledge represented in the two lexicons was actually exploited: the most useful relation is surely *hyperonymy* but also *synonymy*, although to a minor extent. Other relation types were used but not with consistent advantage, like those expressing cross-part-of-speech

synonymy, but also *meronymy* and *holonymy*. A type of connection that seems particularly useful is the one between adjectives and names of locations (cf. 4.3.4) that is instead only partially encoded in ItalWordNet (while it is present in SIMPLE-CLIPS). The most important problem is that it is difficult to insert all the relation types available in the linguistic models of the lexicon into general and systematic strategies to extract information from question and answer. This means that we are able to identify some interesting strategies to exploit the information in the lexicon (Chapter 5) but those strategies are just too granular and specific to be implemented in an Open-Domain application. For example, in SIMPLE-CLIPS we could potentially find some "ready answers" to questions asking about symptoms of diseases (questions like CLEFquestion#98, *Quale è il sintomo del virus Ebola?*, What is a symptom of virus Ebola?). In that lexicon, the relation *Has_As_Effect* is a way to represent a side-effect, a consequence of something. In this sense, it was used to link *morbillo* (measles) and *bolla* (blister), *pneumonia* and *febbre* (fever) etc. I would be possible to concretely encode, when the Answer Type Term is *symptom*, a strategy exploiting the *Has_As_Effect* relation between the specific symptom and the illness in order to individuate a set of possible answers; it would be, however, a very specific strategy. This specific case is an example of a limitation that is not really in the lexicon (where the potentially useful information is available) but rather is in the system, where it is diffucult to foresee such a fined-grained, almost ad-hoc strategy. This case, among the others presented in Chapter 5, shows that also when the information is encoded and consistent, it is not always easy to exploit it: in Open-Domain Question Answering, there is no limit in the topics of the questions and this makes difficult to tailor specific strategies for all the cases that the application has to handle. Obviously, the possibility of exploiting statistical approaches is in this sense stimulating.

Particularly important is moreover the fact that the system is not able to completely exploit the information suggested by the lexicons: we see that, even when we recognized the cases for which methods based on hyperonymy are not adequate (the many cases of Answer Type Terms strongly connoted by a telic, agentive or constituency dimension), still we did not devise alternative strategies of a certain generality able to exploit the information about *telicity*, *agentivity* and *constituency* to drive the search of the answer in a consistent and robust way (by relying on the correspondent semantic relations). The results suggest that if it is immediately feasible to formulate strategies and methods for the part of the lexicon that can be efficaceously described in terms of formal role, it is really difficult to find solutions for the lexical entries whose meaning is constituted by more complex dimensions. This is surely one of the areas where it should be focalized future research.

Another problem concerns the *scope*, the *range of action* of each strategy: as a matter of fact, usually the immediate relational context of the lexical entry under analysis is not enough for the system to carry out a specific task. Moreover, paragraphs that contain an answer formulated with words not present in the query should be retrieved by using a succession of expansions. The analysis of the results shows how often the application has to exploit the inheritance driven by hyperonymy relation to retrieve the sought information. It happens, for example, when the system has to go along the path connecting, in IWN, the ATT *professione* (profession) and the candidate answer *agente segreto* (secret agent) by passing through the intermediate node

lavoratore (employee). The problem rises when we take into consideration levels beyond the first one or a non-specified number of levels, for example during query expansion: in that case, the application that wants to go beyond the first level should be provided with very efficient and precise strategies for "stopping" the expansion when a potentially valid answer is found. This is why expanding the query with words not comprised by the "first", immediate relational context of the lexical entry is so difficult: it is not easy for the application to understand how many times the query has to be expanded and which "direction" has to be followed (keep on ascending through the hyperonym chain or trying the exploitation of the "horizontal" relations as they are encountered during the ascension?). What the system would need is a set of heuristics in order to find useful paths through the lexicon. Computational lexicons can be considered maps where nodes and relations define innumerable paths: heuristics should help the system to "illuminate" only those tracks that, among the others, are meaningful and useful.

Unfortunately, hypothesizing such set of heuristics (in particular those based not only on vertical information) is not easy in Open-Domain Question Answering: the "record of cases" that should be treated and considered in Open-Domain is very high and the strategies that could be adopted have often a granularity coincident with the single synset or SemU (the heuristic to retrieve the symptoms of a disease, the one to retrieve the set of all the professions etc.). For what we observed (Chapter 5), it is not always possible to identify strategies that enable the system to work on wide portions of lexicon and to operate systematic generalisations by univocally exploiting available information. We cannot leave out of consideration the fact that the possibility to rely on systematic and robust information is a fundamental prerequisite for building systems that are effective and efficient, i.e. working effectively with the minimum of effort (a feature no application developer would ever renounce).

We think that one of the conclusions we have reached with this work is that it is not really possible to disjoint the results of the application from the evaluation of the lexicons. This could be considered a negative consequence of the methodology used to evaluate the lexicon. As a matter of fact, as we already pointed out in the introduction, the methodology adopted to carry out our inquiry relies on the results provided by a specific system and the application becomes, in this sense, not a constant but rather a variant of the problem; the design and implementation of the system have a strong impact on the conclusions that we draw. It is obvious that a different system, able to instantiate more "intelligent" and "expert" solutions, might have resolved some of the difficulties we met on our path. This is why it is not always easy to understand if a negative result is obtained because of a limit in the way particular information is represented in our lexicons or in the way the application handles the available information. The importance of such a doubt is *mitigated* by the observation that no state-of-art, existing QA system is able to fully and completely exploit the bulk of information provided by computational lexicons, even if we see that most of the systems uses some information (typically from WordNet) (cf. Chapter 2). We highlighted the word *mitigate* because it is obvious that the doubts connected to the possibility of really evaluating a lexicon and a lexicon model by observing the way it interacts with an application is something that carries some intrinsic problems. In our evaluation, in some way, the application and the representation collapse into each other and become one

single thing, something constituted by dynamic and static components that strongly interact and whose design and management should proceed at the same rate.
In this sense, application and lexicon are completely joined and the limits of the one and of the other tally when the provided representation meets the requirements of the application in terms of usefulness and computability.

We are aware that stating that the lexicon and the application are joined and should be evaluated together is something that can be objected under a theoretic point of view. In (Russel and Norvig, 1998) we read that a language of representation should be expressive, concise, not ambiguous and context-independent. It should be also efficacious: an inference procedure should exist enabling the generation of new inferences based on the formula in the language. Ideally, a strong separation should be kept between the knowledge base and the inference procedure. This should allow the encoder of the KB to worry only about the content of the knowledge and not about the way it will be used in the inference procedure. Assuring efficiency should be a task for the designer of the inference procedure and this aim should not distort the representation the knowledge used by the procedure itself. This kind of vision is just what seems to have driven the research in computational linguistics as far as lexicon and application are concerned. In the years, the two communities of lexicon and application developers have advanced, often without a strong and real collaboration. On one side, who designs, builds and encodes lexicons has aimed to create wide repositories of *multi-purposed*, application-*independent* lexical and semantic knowledge (cf. the definition of knowledge-base by (Amsler, 1884) we reported in the Introduction). On the other side, application developers have always aimed to obtain the best results in terms of effectiveness and efficiency, interacting with lexicons as a black box and trying to obtain from them what they could without knowing them from the inside. Also in (Russel and Norvig, 1998), however, we find a first introduction of the necessity for the developers or the knowledge base and of the inferential engine to work in a more collaborative way (the same idea is also expressed in Calzolari, 2006). As a matter of fact, (Russel and Norvig, 1998) admit that in practice, the knowledge engineer should be aware of how the inference works in order to design the knowledge base to obtain the maximum efficiency. In some way, in this research we have given body to an "alliance" between the lexicon and the application, showing how no "frozen", completely application-independent word meaning exists and that the lexicon should be designed and encoded in such a way to be actually exploitable. The full computability of the lexicon is surely a goal towards which the lexicon designers and encoders should aspire. Nevertheless, as many cases of system failures show, even having at disposal consistent hierarchies and relations, the application does not find in the lexicon everything it needs to succeed. Reasons seem somehow intrinsic: experiments show that the "right" distinction among the senses of a lemma does not exist but rather that the sense continuously changes with almost every context and sub-task of the application (cf. 5.2). In the same way, there seems not to be the "right" synonym on which the system can rely in order to correctly expand a query term, but rather every context seems to bring along not a synonym but a set of paraphrases that probably are in the answer in place of the expression used in the question. Everything seems to suggest that the real, intrinsic limit of lexical meaning as instantiated in language resources is just its being

discrete, de-situated and symbolic. Semantic lexicons are the attempt to entrap something complex and transient (the lexical meaning) in discrete structures defined by following top-down approaches. Those structures are designed mostly indifferently of the way the representation will suit the exigencies of concrete applications that have to analyse the language not in abstract but as it is instantiated in the contexts of the documents. In 4.3.7 we showed how difficult it is to find valid heuristics to go through the many, possible semantic paths that can be traced in a semantic net, aiming at building something similar to the inferential chains presented in (Harabagiu and Moldovan, 1998). Also in (Lin *et al*. 2001) a strong doubt is expressed about the actual possibility to discover inference chains in hand-crafted, static LRs, which presupposes a certain notion of word meaning, i.e. static, relational, discrete, in some way context-independent. For (Lin *et al*. 2001), it is very difficult for humans to encode word meaning with awareness able to built LRs exploitable as basis for sound, robust and effective inference. Lin and Pantel propose an alternative approach based on techniques for the induction of information directly from corpora[89]. Semantic lexicons proved to be really useful in specific modules of the QA system (in particular in determining the expected answer type) while their exploitation has been partly disappointing when applied to query expansion or to answer detection based on relations different from hyperonymy. Surely, the need to capture the reality of the language as instantiated in free texts is something that makes statistical, distributional approaches very alluring, also in view of recovering some of the shortages of semantic lexicons. The most promising idea seems to be boosting the "inferential" potentialities of static, hand-generated LRs with information dynamically acquired from texts in order to fill the gap between question and answer in a more robust, scalable and less-expensive way. The interplay between static lexical information and dynamic information acquired from text via processing is one of the ways computational lexicons could be improved and renewed in the future.

[89]Lin and Pantel's methodology broadens the scope of Harris' Distributional Hypothesis from the word to the dependency trees of parsed corpus.

6 Bibliography

Abney S., Collins M., and A. Singhal. *Answer extraction*. In Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000), 2000.

Alonge A., Calzolari N., Vossen P., Bloksma L., Castellon I., Marti T., and Peters W., *The Linguistic Design of the EuroWordNet Database*. In N. Ide *et al*. (eds.) Computers and the Humanities, Special Issue on EuroWordNet, 32(2-3), Kluwer Academic Publishers, Dordrecht, 1998.

Amsler R. A., *Lexical knowledge bases*. In Proceedings of the 10th international conference on Computational linguistics, Stanford, California, 1984.

Amsler R., A., *Research Toward the Development of a Lexical Knowledge Base for Natural Language Processing*. In Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York, NY, USA , 1989.

Amsler R.A., *Words and Worlds*. In Proceedings of the Third Workshop on theoretical Issues in Natural Language Processing (TINLAP-3), Las Cruces, NM, 16-19, 1987.

Apresjan J., *Regular Polysemy*. In Linguistics, 142, pp. 5-32, 1973.

Attardi G., Cisternino A., Formica F., Simi M., Tommasi A., and Zavattari C., *PIQAsso: Pisa Question answering System*. In Proceeding of the 10th TREC Conference, 2001.

Attardi G., Cisternino A., *Reflection support by means of template metaprogramming*. In Proceedings of Third International Conference on Generative and Component-Based Software Engineering, LNCS, Springer-Verlag, Berlin, 2001.

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

Bartolini R., Lenci A., Montemagni S., and Pirrelli V., *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, 2002.

Battista M and Pirrelli V., *Una Piattaforma di Morfologia Computazionale per l'Analisi e la Generazione delle Parole Italiane*, ILC-CNR Technical Report, 1999.

Berger A., Caruana R., Cohn D., Freitag D. and Mittal V., *Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding*. Research and Development in Information Retrieval, pp. 192-199, 2000.

Berland M. and Charniak E., *Finding Parts in Very Large Corpora*. In ACL-1999. pp. 57.64. College Park, MD, 1999.

Bertagna F., Chiran L., Simi M., *ILC-UniPi Italian QA*. In Peters C. and Clough P. and Gonzalo J. and Jones G. and Kluck M. and Magnini B. (eds.), Fifth Workshop of the Cross--Language Evaluation Forum (CLEF-2004), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2005.

Bertagna F., *Italian Language Resources in a Question Answering Task*. In Proceedings of the 2nd

CoLogNET-ElsNET Symposium dedicated to "Questions and Answers: Theoretical and Applied Perspectives", Amsterdam, The Netherlands, 2003.

Bertagna F., *Using Semantic Language Resources to Support Textual Inference for Question Answering*. In Proceedings of the Fourth LREC Conference, 2004.

Bertuccelli Papi M., *Implicitness in Text and Discourse*, Edizioni ETS, Pisa, 2000.

Bilotti M., Katz B., and Lin L., *What Works Better for Question Answering: Stemming or Morphological Query Expansion?*. In Proceedings of the Information Retrieval for Question Answering (IR4QA), Workshop at SIGIR 2004, Sheffield, England, 2004.

Boguraev B.K., Briscoe T., *Large Lexicons for Natural Language Processing*. Computational Linguistics, 13(3-4), Special issue of the lexicon, pp. 203-218, 1987.

Boguraev B.K., *The Definitional power of words*. In Proceedings of the Third Workshop on theoretical Issues in Natural Language Processing (TINLAP-3), Las Cruces, NM, pp. 11-15, 1987.

Bransford J. D., Sharp D. M., Vye N. J., Goldman S. R., Hasselbring T. S., Goin L., O'Banion K., Livernois J., Saul E., *MOST environments for accelerating literacy development*. In S. Vosniadou, E. De Corte, R. Glaser, & H. Mandl (eds.), International perspectives on the psychological foundations of technology-supported learning environments, pp. 223-256, Hillsdale, NJ: Erlbaum, 1996.

Breck, E.J., Burger J.D., Ferro L., Hirschman L., House D., Light M., Mani I., *To Evaluate Your Question Answering System Every Day …and Still Get Real Work Done*. Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, pp. 1495-1500, 2000.

Bronnenberg W., Bunt H., Landsbergen S., Scha R., Schoenmakers W., Van Utteren E., *The question answering system PHLIQA1*. In L. Bolc (ed.), Natural Language Question Answering Systems, pp. 217-305, MacMillan, 1980.

Buguraev B., Levin B., *Models for Lexical Knowledge Bases*. In J. Pustejovsky (ed.) Semantics and the Lexicon, Kluwer Academic Publisher, the Netherlands, 1993.

Burger, J., C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C. Y. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, E. Voorhees, R. Weishedel, *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)*, 2001. Available at: http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc.

Busa F., Bouillon P. (eds.), *The Language of Word Meaning*, Cambridge University Press, UK, 2001.

Busa F., Calzolari N., Lenci A., *Generative Lexicon and the SIMPLE-CLIPS model*. In P. Bouillon and F. Busa (eds.), The Language of Word Meaning, Cambridge University Press, 2001.

Byrd R.J., Calzolari N., Chodorow M.S., Klavans J.L., Neff M.S., Rizk O.A, *Tools and Methods for Computational Lexicology*. In Computational Linguistics 13(3-4), pp. 219-240, 1987.

Calzolari N., *Technical and Strategic issues on Language Resources for a Research Infrastructure*. In S. Furui (eds.) Proceedings of the International Symposium on Large-scale Knowledge Resources (LKR2006), Tokyo Institute of Technology, pp. 53-58, 2006.

Calzolari N., *Acquiring and representing semantic information in a lexical knowledge base*. In Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation, Berkeley, California, pp. 188-197, 1991.

Calzolari N., Bertagna F., Lenci A., Monachini M. (eds.). *Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entries)*, ISLE Deliverable 2.2 &2.3 CLWG, Pisa, 2002. Available at http://lingue.ilc.cnr.it/EAGLES96/isle/ .

Calzolari N., D*etecting Patterns In A Lexical Data Base*. In Proceedings of the 22nd Annual Meeting of the ACL, 1984.

Calzolari N., Hagman J., Marinai E., Montemagni E., Spanu A., Zampolli A., *Encoding Lexicographic Definitions as Typed Feature Structures*. In R. Posner and G. Meggle (eds.), Theorie un Praxis des Lexikons, Walter de Gruyter, Berlin, 1993.

Calzolari N., Soria C., Bertagna F., Barsotti F., *Evaluating lexical resources using Senseval*. In Journal of Natural Language Engeneering, Special Issue on Senseval-2, 8(4), pp. 375-390, 2002.

Calzolari N., *The dictionary and the thesaurus can be combined*. In Walton Evens M. (ed.), Relational Models of the lexicon, Cambridge University Press, GB, pp. 75-95, 1988.

Calzolari, N., Grishman, R., Palmer, M. (eds.), *Survey of major approaches towards Bilingual/Multilingual Lexicons*. ISLE Deliverable D2.1-D3.1, Pisa, 2001.

Caramazza A., Shapiro K., *Language categories in the brain: Evidence from aphasia*. In L. Rizzi and A. Belletti (eds.), Structures and Beyond. Oxford University Press, Oxford, UK, in press.

Chodorow, M.S., R.J. Byrd, G.E. Heidorn, *Extracting semantic hierarchies from a large on-line dictionary*. In Proceedings of the 23rd Annual ACL Conference, Chicago, Ill., pp. 299-304, 1985.

Clarke C., Cormack G. and Lyman T., *Exploiting redundancy in question answering*. In Proceedings of SIGIR'2001, 2001.

Clifton T., Colquhoun A., Teahan W., *Bangor at TREC 2003: Q&A and Genomics Tracks*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Cohen P., Schrag R., Jones E., Pease A., Lin A., Starr B., Gunning D., Burke M., *The DARPA High Performance Knowledge Bases Project*. In AI Magazine, 18(4), pp. 25-49, 1998 (available at http://www.findarticles.com/p/articles/mi_m2483/is_4_19/ai_53560912/pg_1)

Cruse D.A., *Lexical Semantics*, Cambridge University Press, Cambridge, 1986.

Dagan I. and Glickman O., *Probabilistic textual entailment: Generic applied modeling of language variability*. In Learning Methods for Text Understanding and Mining, Grenoble, France, 2004.

De Mauro T. (ed), *Il Dizionario della Lingua Italiana per il Terzo Millennio*, Paravia, 2000.

Devitt A. and Vogel C., *The Topology of WordNet: Some Metrics*. in P. Sojka, K. Pala, P. Smr, C. Fellbaum, P. Vossen (eds.): GWC 2004, Proceedings, pp. 106-111, Masaryk University, Brno, 2003.

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

Echihabi A., Hermjakob U., Hovy E., Marcu D., Melz E., Ravichandran D., *Multiple-Engine Question Answering in TextMap*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Erbach G., *Evaluating Human Question answering Performance under Time Constraints*. In Working Notes for the CLEF 2004 Workshop, pp. 15-17 September, Bath, Uk, 2004.

Fano R., *Transmissions of Information: A Statistical Theory of Communications*, MIT Press, 1961.

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

Ferro L., *TREC Answer Key*, Questions 1-200, 1999. Available at http://trec.nist.gov/data/qa/add_QAresources/trec8_answerkey.txt

Fillmore, Charles J., *An alternative to checklist theories of meaning*. Papers from the 1st Annual Meeting of the Berkeley Linguistic Society, pp. 123-132, 1975.

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

Gaizauskas R., Greenwood M.A., Hepple M., Roberts I., Saggion H., and Sargaison M., *The University of Sheffield's TREC 2003 Q&A Experiments*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Gale W., Church K., and Yarowsky D., *Estimating upper and lower bounds on the performance of word-sense disambiguation programs*. Proceedings of the 30th annual meeting on Association for Computational Linguistics, 1992.

Gangemi A., Guarino N., Masolo C. and C. Oltramari A., *Understanding Top-Level Ontological Distinctions*. In Proceedings of IJCAI 2001 Workshop on Ontologies and Information Sharing, 2001 (2001a).

Gangemi A., Guarino N., Oltramari A., *Conceptual Analysis of Lexical Taxonomies: the Case of WordNet Top-Level*. In C. Welty, B. Smith (eds.), Proceedings of the 2001 Conference on Formal Ontology and Information Systems, Amsterdam, IOS Press, 2001 (2001b).

Godfrey J.J., Zampolli A., *Overview of the Chapter on Language Resources*. In R. Cole *et al*. (eds.) Survey of the State of the Art in Human Language Technology, Giardini Editori, Pisa, 1997.

Goldman, S. R., *Inferential reasoning in and about narrative texts*. In A. Graesser & J. Black (eds.), The psychology of questions, pp. 247-276, Hillsdale, NJ, Erlbaum, 1985.

Gonzalo J., Verdejo F., Peters C., Calzolari N., *Applying EuroWordNet to Cross-Language Text Retrieval*. In N. Ide *et al*. (eds.) Computers and the Humanities, Special Issue on EuroWordNet, 32(2-3), Kluwer Academic Publishers, Dordrecht, 1998.

Graesser A.C., Murachver T., *Symbolic procedures of question answering*. In A.C. Graesser, J.B. Black (eds.), The psychology of questions. Lawrence Hillsdale, NJ, Erlbaum Associates, pp. 15-88, 1985.

Graesser A.C., Olde B., Pomeroy V., Whitten S., Lu S., Craig S., *Inferences and Questions in Science Text Comprehension*. In Otero, J., Leon, J.A., & Graesser, A.C.(eds.) The psychology of science text comprehension, Mahwah, NJ, Erlbaum, 2002.

Graesser, A.C., Singer, M., & Trabasso, T., *Constructing inferences during narrative text comprehension*. In Psychological Review, 101, pp. 371-395, 1994.

Graesser, A.C., Wiemer-Hastings, P., & Wiemer-Hastings, K., *Constructing inferences and relations during text comprehension*. In T. Sanders, J. Schilperoord, & W. Spooren (eds.), Text representation: Linguistic and psycholinguistic aspects, pp. 249-271, Amsterdam, Benjamins, 2001.

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

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

Grunfeld L., Kwok K.L., Dinstl N., Deng P., *TREC 2003 Robust, HARD and QA Track Experiments using PIRCS*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Guarino N., *Some Ontological Principles for Designing Upper Level Lexical Resources*. In First International Conference on Language Resources and Evaluation Granada, Spain, 28-30 May 1998.

Guazzini E., Ulivieri M., Bertagna F., Calzolari N., *Senseval-3: the Italian All-Words Task*. In SENSEVAL-3 proceedings, 2004.

Hanks P., *Do word meanings exist?*. In A. Kilgarriff and M. Palmer (eds.), Special Issue on Senseval. Computers and the Humanities 34(1-2), pp. 205-215, 2000.

Harabagiu S. and Moldovan D., *Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text*. In S. Shapiro and L. Iwanska (eds), Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language AAAI/MIT Press, pp. 301-334, 2000.

Harabagiu S. and Moldovan D., *Knowledge Processing on Extended WordNet*. In C. Fellbaum (ed.) WordNet: An Electronic Lexical Database and Some of its Applications, pp. 379-405, MIT Press, 1998.

Harabagiu S., Moldovan D., Clark C., Bowden M., Williams J., Bensley J., *Answer Mining by Combining Extraction Techniques with Abductive Reasoning*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Harabagiu S., Moldovan D., Pasca M, Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus R. and Morarescu P., *FALCON: Boosting Knowledge for Answer Engines*. In Proceedings of the Text Retrieval Conference (TREC-9), 2000.

Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V. and Morarescu P., *The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering*. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001), pp. 274-28, 2001.

Hearst M., *Automatic acquisition of hyponyms from large text corpora*. In COLING-92. pp. 539-545. Nantes, France, 1992.

Hirschman L., Gaizauskas R., *Natural Language Question Answering: The View form Here*. In J. Tait,

B. K. Boguraev, C. Jacquemin (eds.), L. Hirschman and R. Gaizauskas (Guest Eds.), Natural Language Engineering, Special Issue on Question Answering, 7(4), 2001.

Hovy H., Gerber L., Hermjakob U., Lin C., Ravichandran D., *Towards Semantic-Based Answer Pinpointing*. In Proceedings of Human Language Technologies Conference, pp. 339-345, San Diego CA, 2001.

Ide N. and Veronis J., *Extracting Knowledge bases from machine readable dictionary: have we wasted our time?*. In Proceedings of the International Workshop on the Future of Lexical Research, Beijing, China, pp. 137-46, 1993.

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

Ide N., Wilks Y., *Making Sense About Sense.* In Agirre, E., Edmonds, P. (eds.), Word Sense Disambiguation: Algorithms and Applications, Springer, forthcoming.

Il Grande Dizionario di Italiano con CD-Rom 2006 N.E., Garzanti, 2005.

Jackendoff R., *Foundations of Language. Brain, Meaning, Grammar, Evolution*, Oxford University Press, GB, 2002.

Jijkoun V., Mishne G., Monz C., de Rijke M., Schlobach S., Tsur O., *The University of Amsterdam at the TREC 2003 Question Answering Track*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

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

Katz B., Lin J., Loreto D., Hildebrandt W., Bilotti M., Felshin S., Fernandes A., Marton G., Mora F., *Integrating Web-based and Corpus-based Techniques for Question Answering*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Kilgarriff A., *How dominant is the commonest sense of a word?*. In Sojka, Kopecek and Pala (eds.) Text, Speech, Dialogue. Lecture Notes in Artificial Intelligence, Vol. 3206., Springer Verlag, pp. 103-112, 2004.

Kilgarriff A., *I don't believe in word sense*. In Computers and the Humanities, 31(2), pp. 91-113, 1997.

Kouylekov M., Magnini B., Negri M., Tanev H., *ITC-irst at TREC 2003: the DIOGENE QA System*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Krovetz R. and Croft W.B., *Lexical Ambiguity and Information Retrieval*, ACM Transaction and Information System, 10 (2), pp. 115-141, 1991.

Kupiec J., *Murax: A robust linguistic approach for question answering*. In Proc. 16th Int'l Conference on R&D in IR (SIGIR), pp. 181-190, 1993.

Kwok K.L., Grunfeld L., Dinstl N. and Chan M., *TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS*. Proceedings of the Text Retrieval Conference (TREC-9), pp. 26-35, 2000.

Lee D.L., Chuang H. and Seamons K.., *Document ranking and the vector-space model*. IEEE Software, 14(2), pp. 67-75, 1997.

Leech G., *Being precise about lexical vagueness*, York Papers in Linguistics, 6, 1-31, 1976.

Lehnert W., *The Process of Question Answering*, Lawrence Erlbaum Associated, N.J., 1978.

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

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

Lenci A., Montemagni S., Pirrelli V., Soria C., *FAME: a Functional Annotated Meta-Schema for multi-modal and multilingual Parsing Evaluation*, Proceeding of LREC-2000, 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., *SIMPLE: A General Framework for the Development of Multilingual Lexicons*. In International Journal of Lexicography, 13 (4), pp. 249-263, 2000.

Lewis D.D., *Representation and Learning in Information Retrieval*, doctoral dissertation, Univ. of Massachusetts, Amherst, Mass., 1991.

Lin D., Pantel P., *Discovery of Inference Rules for Question Answering*. In Natural Language Engineering 7(4), pp. 343-360, 2001.

Litkowski, K. C., *Question-Answering Using Semantic Relation Triples*. In Voorhees, E. M. and Harman, D. K. (eds.) Information Technology: The Eighth Text REtrieval Conferenence (TREC-8), NIST Special Publication 500-246. Gaithersburg, MD: National Institute of Standards and Technology, pp. 349-56., 2001.

Litkowski, K. C., *Syntactic Clues and Lexical Resources in Question-Answering*. In Voorhees, E. M. and Harman, D. K. (eds) Information Technology: The Ninth Text REtrieval Conferenence (TREC-9), NIST Special Publication 500-249. Gaithersburg, MD: National Institute of Standards and Technology, pp. 157-66, 2001.

Llopis F., Ferr´ andez A., and Vicedo J., *Passage selection to improve question answering*. In Proceedings of the COLING 2002 Workshop on Multilingual Summarization and Question Answering, 2002.

Lyons J., *Semantics*, Cambridge University Press, London, 1977.

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

Magnini B and Cavaglià G., *Integrating Subject Field Codes into WordNet*. In Gavilidou M., Crayannis G., Makantonau M., Piperidis S., and Stainhoaver G., (eds.), Proceedings of LREC-2000, Second International conference on Language Resources and Evaluation, Athens, Greece, 2000.

Magnini B. and Prevete R., *Exploiting Lexical Expansions and Boolean Compositions for Web*

Querying. In ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong-Kong University of Science and Technology, October 2000.

Magnini B., Negri M., Prevete R., Tavev H., *Multilingual Question/Answering: the Diogene System*. In Proceeding of the 10th TREC Conference, 2001.

Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., de Rijke M., Rocha P., Simov K., Sutcliff R., *Overview of the CLEF 2004 Multilingual Question Answering Track*. In Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, Uk, 2004.

Mandala R., Takenobu T. and Hozumi T., *The Use of WordNet in Information Retrieval*. Proceedings of Coling-ACL, 1998.

Mann G. S., *Fine-Grained Proper Noun Ontologies for Question Answering*. SemaNet. 02: Building and Using Semantic Networks, Taipei, Taiwan, 2002.

Marinelli R., Biagini L., Bindi R., Goggi S., Monachini M., Orsolini P., Picchi E., Rossi S., Calzolari N., Zampolli A., *The Italian PAROLE corpus: an overview*. In Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa, Special Issue of Linguistica Computazionale, Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-roma, 2003.

Massot M., Rodriguez H., Ferres D., *QA UdG-UPC System at TREC-12*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Maybury M.T., Sparck Jones K.,Voorhees E., Harabagiu S., Liddy L., Prange J., *Workshop on Strategy and Resources for Question Answering in conjunction with the Third International Conference on Language Resources and Evaluation (LREC)*. Palacio de Congreso de Canarias,Canary Islands,Spain, 2002.

Maybury M.T., *Toward a Question Answering Roadmap*, 2002. Available at http://www.mitre.org/work/tech_papers/tech_papers_02/maybury_toward/

Mihalcea R., Chklovski T., Kilgarriff A., *The Senseval-3 English Lexical Sample Task*. In Proc. Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona, July, pp. 25-28, 2004.

Mihalcea R., *The Semantic Wildcard*. In Proceedings of the LREC 2002 Workshop on "Using Semantics for Information Retrieval and Filtering: State of the Art and Future Research", Las Palmas, Spain, May 2002.

Miller G. A., *WordNet: a dictionary browser*. In Proceedings of the First International Conference on Information in Data, University of Waterloo, Waterloo, 1985.

Miller G., Beckwith R., Fellbaum C., Gross D., Miller K.J., *Introduction to WordNet: An On-line Lexical Database*. In International Journal of Lexicography, 3(4), pp. 235-244, 1990.

Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R. and Rus V., *The Structure and Performance of an Open-Domain Question-Answering System*. In Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL-2000), Kong Kong, October 2000, pp. 563-570, 2000.

Monarch I. and Carbonell J., *CoalSORT : A Knowledge-Based Interface*, IEEE Expert, pp. 39 - 53, 1987.

Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Lenci A., Pirrelli V., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M. T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R., *The syntactic-semantic Treebank of Italian. An Overview.* In Linguistica Computazionale a Pisa vol. I, pp. 461-492, 2003.

Monz C., *Document Retrieval in the Context of Question Answering*. In F. Sebastiani (ed.) Proceedings of the 25th European Conference on Information Retrieval Research (ECIR-03), Lecture Notes in Computer Science 2633, Sprinter, pp. 571-579, 2003 (2003a)

Monz C., *From document Retrieval to Question Answering*, PhD thesis, ILLC, U. of Amsterdam, 2003. (2003b)

Nirenburg S, Raskin V., *Ten Choices in Lexical Semantics*, MCCS-96-304, CRL, NMSU, 1996.

Nirenburg S., Wilks Y., *What's in a Symbol: ontology, representation and language*. Journal of Theoretical and Experimental AI (JETAI), 1999.

Nyberg E., Mitamura T., Callan J., Carbonnell J., Frederking R., Collins-Thompson K., Hiyakumoto L., Huang Y., Huttenhower C., Judy S., Ko J., Kupsc A., Lita L.V., Pedro V., Svoboda D., and Van Durme B., *The JAVELIN Question-Answering System at TREC 2003: A Multi-Strategh Approach with Dynamic Planning*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Palmer M., Dang H. and Fellbaum C., *Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically*. In Journal of Natural Language Engineering, 2001.

Paranjpe D., Ramakrishnan G., Srinivasan S., *Passage Scoring for Question Answering via Bayesian Inference on Lexical Relations*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Paşca M, Harabagiu S., *High Performance Question/Answering*. In Proceedings of the 24th Annual International ACL SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2001), September 2001, New Orleans LA, pp. 366-374, 2001.

Paşca M., *Open-Domain Question Answering from Large Text Collections*, CSLI Studies in Computational Linguistics, 2003.

Pazienza M. T., Pennacchiotti M., Zanzotto F. M., *Textual Entailment as Syntactic Graph Distance: a rule based and a SVM based approach*. In Proceedings of the PASCAL Challenges Workshop, Southampton, U.K., April 2005.

Peters W., Peters I. and Vossen P., *Automatic Sense Clustering in EuroWordNet*. In Proceedings of the First Interational Conference on Language Resources and Evaluation Granada, 1998.

Phillips A.V., *A question-answering routine*. In Memo 16, Artificial Intelligence Project, MIT, Cambridge, Mass., 1960.

Picchi E., D.B.T., *A Textual Data Base System*. In Computational Lexicology and Lexicography. In Special Issue dedicated to Bernard Quemada, Linguistica Computazionale, Pisa, 1991.

Poesio M., Ishikawa T., Schulte im Walde S., Vieira R., *Acquiring Lexical Knowledge for Anaphora Resolution*. In Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC), 2002.

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

Prager, J., Brown, E., Coden, A. & Radev, D., *Question-answering by predictive annotation*. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 184-191, 2000.

Pulman S.G., *Word Meaning and Belief*, Croom Helm, London, 1983.

Pustejovsky J., *The Generative Lexicon,* MIT Press, Cambridge MA., 1995.

Rada R., Mili H., Bicknell E., and Blettner M., *Development and Application of a Metric on Semantic Nets*. In IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 1, pp. 17-30, 1989.

Radev R., Prager J., and Samn V., *Ranking suspected answers to natural language questions using predictive annotation*. In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-2000), pages 150–157, Seattle,Washington, 2000.

Raphael B.: *SIR: A computer program for semantic information retrieval*. MAC-TR2 Project MAC MIT, 1964.

Renzi L., Salvi G., Cardinaletti A. (eds.), *Grande grammatica italiana di consultazione*, Bologna, Il Mulino, 1995.

Richardson R., Smeaton A.F., *Using WordNet in a Knowledge-based approach to Information Retrieval*, School of Computers Applications working Paper CA-0395, 1995.

Riloff, E. and Shepherd, J., *A corpus-based approach for building semantic lexicons*. In Proceedings of EMNLP-1997, 1997.

Rinaldi F., Dowdall J., Hess M., Mollá D., Schwitter R., *Question Answering in Terminology-rich Technical Domains*. In New Directions in Question Answering (collection edited by Mark Maybury), AAAI press, 2003

Robertson S.P., Weber K., Ullman J.D., Mehta A., *Parallel Question Parsing and Memory Retrieval*. In Journal of Memory and Language, 32, pp. 155-168, 1993.

Rodriguez H., Climent S., Vossen P., Bloksma L., Roventini A., Bertagna F., Alonge A., Peters W., *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*. In N. Ide *et al*. (eds.) Computers and the Humanities, Special Issue on EuroWordNet, 32(2-3), Kluwer Academic Publishers, Dordrecht, 1998.

Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A., *ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian*. In Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa, Special Issue of Linguistica Computazionale, Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-Roma, 2003.

Roventini A., Ulivieri M., Calzolari N., *Integrating Two Semantic Lexicons, SIMPLE-CLIPS And ItalWordNet: What Can We Gain?*. In Proceedings of the LREC-2002 Conference, Las Palmas di Gran Canaria, Spain, 2002.

Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M.C., Ulivieri M., Rossi S., *A Computational Semantic Lexicon of Italian: SIMPLE*. In Zampolli A., Calzolari N., Cignoni L.

(eds.), Computational Linguistics in Pisa, Special Issue of Linguistica Computazionale, Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-Roma , 2003.

Russell S., Norvig P., *Intelligenza Artficiale: un approccio moderno*, Torino, UTET, 1998.

Sabatini F., Coletti V., *DIZIONARIO DISC. Dizionario di italiano*, Roma, Giunti, 1996.

Sanfilippo A., Calzolari N., Ananiadou S., Gaiazauskas R., Saint-Dizier P., Vossen P., Alonge A., Bel N., Bontcheva K., Bouillon P., *EAGLES Recommendations on Semantic Encoding*, 1999. Available at http://www.ilc.pi.cnr.it/EAGLES96/rep2

Schütze H., Pedersen J., *Information Retrieval based on word senses*. In Proceedings of the Fourth Annual Symposyum on Document Analysis and Information Retrieval, Las Vegas, pp. 161-175, 1995.

Simmons R. F., *Answering English Questions by computers: A Survey*. In Communications of the ACM, 8(1), pp. 53-70, 1965.

Simmons R.F. and McConlogue K.L., *Maximum-depth indexing for computer retrieval of English language data*. In American Documentation, 14, 1, pp. 68-73, 1963.

Snyder B. and Palmer M., *The English all-words task*. In Proceedings of SENSEVAL-3, Barcelona, Spain, 2004.

Soria C., Tesconi M., Bertagna F., Calzolari N., Marchetti A., and Monachini M., *Moving to Dynamic Computational Lexicons with LeXFlow*. Accepted for publication in Proceedings of LREC2006, Genova, Italy, 2006.

Sparck Jones K., *What is the role of NLP in Text Retrieval?*. In Strzalkowski T. (ed.), Natural Language Information Retrieval, Kluwer, 1999.

Srihari R. and Li W., *A Question Answering System Supported by Information Extraction*. In Proceedings of the 1 st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL-00), pp. 166-172, 2000.

Stone P.J., Bayles R.F., Namerwirth J.Z, Ogilvie D.M., *The General Inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information*. In Behavioral Science, 7, 4, pp. 1-15, 1962.

Thorne J.P., *Automatic Language Analysis*. ASTIA 297381, Final Tech. Rep., Arlington, Va., 1962.

Voorhees E. M., *Query Expansion Using Lexical-Semantic Relations*. In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, pp. 61-69, 1994.

Voorhers E.M., *The TREC-1999 Question Answering Track Report*. In Proceedings of the 8th Text REtrieval Conference (TREC-8), pp. 77-82. Gaithersburg, Maryland, NIST, 1999.

Vossen, P. (ed.), *EuroWordNet General Document*, 1999. Available at http://www.hum.uva.nl/~ewn.

Vossen, P., *Introduction to EuroWordNet*. In N. Ide *et al*. (eds.) Computers and the Humanities, Special Issue on EuroWordNet, 32(2-3), Kluwer Academic Publishers, Dordrecht, 1998.

Weizenbaum J., *ELIZA – A computer program for the study of natural language communication between man and machine*, Communication of the ACM, 9 (1), pp. 36-45, 1966.

Wendlandt E. and Driscoll J., *Incorporating a Semantic Analysis into a Document Retrieval Strategy*, ACM SIGIR, 1991.

Wilks Y., *Stone Soup and the French Room*. In Zampolli A., Calzolari N. and Palmer M. (eds.), Current Issues in Computational Linguistics: In honor of Don Walker. Pisa, Italy: Giardini and Dordrecht, The Netherlands, Kluwer, pp. 585-595, 1994.

Winograd T., *Understanding Natural Language*. Academic Press, 1972.

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

Woods W., *Progress in Natural Language Understanding – an Application to Lunar Geology*. In AFIPS Conference Proceedings, 1973.

Wu L., Huang X., Zhou Y., Du Y., You L., *FDUQA on TREC2003 QA task*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Xu J., Licuanan A., Weischedel R., *TREC 2003 QA at BBN: Answering Definitional Questions*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Yang H., Cui H., Maslennikov M., Qiu L., Kan M.-Y., Chua T., *QUALIFIER In TREC-12 QA Main Task*. In Proceedings of the Twelfth Text Retrieval Conference, 2003.

Zue V., Seneff S., Glass J., Polifroni J., Pao C., Hazen T.J., Hetherington L., *Jupiter: A Telephone-Based Conversational Interface for Weather Information*. In IEEE Transactions on Speech and Audio Processing, 2000.