*Artificial Intelligence and Law* **8:** 1–34, 2000.



# Natural language processing for transparent communication between public administration and citizens

BERNARDO MAGNINI, ELENA NOT, OLIVIERO STOCK and CARLO STRAPPARAVA
*IRST, 38050 Povo, Trento, Italy*
*E-mail: {magnini,not,stock,strapparava}@irst.itc.it*

**Abstract.** This paper presents two projects concerned with the application of natural language processing technology for improving communication between Public Administration and citizens. The first project, GIST, is concerned with automatic multilingual generation of instructional texts for form-filling. The second project, TAMIC, aims at providing an interface for interactive access to information, centered on natural language processing and supposed to be used by the clerk but with the active participation of the citizen.

**Key words:** dialogue systems, multilingual NL generation, natural language processing, public administration

## 1. Introduction

*Transparency* has become a key request to public administrations. The concept has a general 'objective' meaning (i.e., things are not hidden or veiled) and a more communication-oriented meaning. From the first point of view information should be comprehensible, timely, should concern general issues and issues relevant for the specific individual, should be explorable and be put in context. From the second point of view transparency means that information should be conveyed in a modality that favours a more direct and collaborative relationship between citizens and Public Administration. Indeed transparency has implications for democracy, legal considerations and efficiency of the Public Administration-citizens system.

The need for communication with the citizen may occur for different reasons, including:

- to provide detailed information on the progress of an administrative procedure or on the services available;
- to provide clear instructions on the obligations the citizen must fulfil and on the documents they have to submit;
- get punctual information on the personal and family situation of the citizen.

In any case, the fundamental requisite for a *transparent* communication, is that information is available and reaches every citizen with the same clarity, to avoid disparity.

Networks, integration of data bases and specific access programs are at the heart of a lot of activity in all advanced countries. This implies that data are becoming available independently of the location where they are physically stored and in principle they can be seen by all. The real problem, still underestimated at least from a technological point of view, is communication with the citizen, where the issue is not just a matter of connectivity. We think natural language processing technology can help by proposing concrete and realistic solutions.

In this paper we shall present two projects we have been working on recently (both were sponsored by the European Commission under the Language Research and Engineering programme and derivatives). They deal with two different aspects of transparency, particularly relevant in our society. The first, GIST (described in Sections 1–4), deals with the automatic multilingual production of texts for the citizen. Multilinguality is an important aspect of European societies, both internally (many countries have minorities that enjoy some linguistic rights) and globally (Europe will develop as an officially multilingual entity). Other than the strictly legal commitments to multilingual communication, there will be a growing need to take into account the communication needs of immigrants and foreigners. The domain of the GIST project is instructional text for form completion, a theme that is *per se* (even for monolingual applications) of substantial impact.

The second project (its description starts in Section 5), called TAMIC, deals with information access to multiple information for the citizen. It envisages a natural language centered interface to be shared by clerk and citizen in the quest for specific information. The idea is that, to obtain transparency, the role of a human clerk is normally essential: for most citizens, if the information is not trivial, human assistance is of great value. Yet, with many different data bases available from one window, you need an interface that can be used by an employee who may be expert only on some of the information available. Furthermore the citizen should understand all the ongoing communication between clerk and system and participate with the help of the assistant. So, natural language dialogue capabilities, integrated with a modern view of the interface, become a very promising concept, just for its inherent aspects of a natural means of communication.

Both projects involved an analysis of users' needs and work with potential users of the system, and both projects have followed a concrete, constructive approach. Yet, of course, their results should be seen only at the level of advanced prototypes.

The descriptions will leave aside technicalities but convey the essence of the projects.

## 2. Helping the Citizen to Complete Legal Requirements: Producing Transparent Forms

Application forms represent one of the main communication channels between the citizens and Public Administration. Whenever citizens want to apply for a document or a benefit, they are required to fill out a form, specifying various kinds of information, from personal details to income data. The requested information reflects the current status of the legislation and may be quite complex. The correct and complete form content is often essential for the legal procedure to start. For this reason, forms very often include instructions that help the applicant fill them out.

Producing clear and effective application forms, is a major and permanent effort for large public institutions. Every time some changes are introduced in the current legislation about the services offered to citizens or the obligations expected from them, new application forms need to be created or old ones need to be revised. The problem is even more complex in multilingual areas, where public documentation must appear in all the official languages. Administrative agreements between different countries (for example in the pension domain) are another source of multilingual forms, and so in the future could massive immigration.

Many public administrations take advantage of information technology tools to simplify and speed up the form production cycle: text editors augmented with on-line dictionaries or grammar checkers are the tools most widely used by administrative experts and technical writers, but also tools for graphic layout arrangement (once used only at the end of the specification process by graphic designers or typographers) are used more and more throughout the form creation phase. Advances in the Artificial Intelligence field may provide valuable help for the development of innovative tools, relieving administrative experts and translators of part of the burden involved in form production. In particular, the results in Automatic Generation of Natural Language (Generation for short) and Machine Translation are the most promising and are already being used in real applications.

In this section of the paper we will focus on the advantages that automatic (multilingual) generation of natural language can bring to the form production cycle.

### 2.1. REQUIREMENTS FOR GOOD FORMS

“It is the writer’s responsibility to be clear. It is NOT the reader’s responsibility to understand” [Prof. R. Eagleson, *Writing in plain English* 1990].

In the past, technical writers often had no source of information other than their common sense and experience in how to write good administrative and legal texts. No special courses were organised and no professional guide books were available. The administrative language was considered, as legal language, a sort of language apart from every day life, with its own words and rules for formal linguistic expression. With the emerging need for transparency, technical writers are now urged to be effective communicators and to convey their message in clear and concise terms, and the problem of establishing common guidelines for good quality documents has become a critical one. Efforts towards this goal have been made in parallel in different countries where special committees have been organised to study the simplification of the bureaucratic procedures and communication channels with the citizens. The guidelines for transparency that emerged contain general principles of good writing, such as:

- Don't use unnecessary words.
- Keep sentences short (about 20/30 words).
- Define new terms and concepts.
- State obligations clearly, avoiding ambiguities and misunderstandings.

But guidelines also include detailed suggestions for the simplification of grammatical and lexical structures, such as:

- Limit the use of nominalizations, impersonal forms and passive constructions that help increase the social distance between writer and reader.
- Avoid the use of subjunctive that could give rise to complex grammatical constructions.
- Be consistent with terminology; make sure the definition of a term does not change throughout the text.

For writing forms, in particular, all the standard technical writing guidelines apply but many other features (like layout and helping notes) have to be taken into consideration to encourage users to read the forms and fill them out. In 1983, the British Department for Social Security published the Good Forms Guide, the seminal reference book for all technical writers of British government forms. This guide contains many detailed suggestions on how to write and structure good forms to be distributed to a large range of people. Since its publication, a substantial improvement in the quality of the forms has been obtained and, most importantly, market research on sets of target readers have been conducted to validate the readability and clearness of forms before they are printed and distributed.

### 2.2. EQUAL OPPORTUNITIES FOR CITIZENS IN A MULTILINGUAL COMMUNITY

By law, in multilingual areas all the documentation that has to be distributed to the public has to be available in all the official languages to guarantee equal opportunities also to linguistic minorities. An example of multilingual application scenario is the Italian bilingual province of Bolzano. Currently, all the public documentation that circulates inside the province is first produced in Italian and then translated

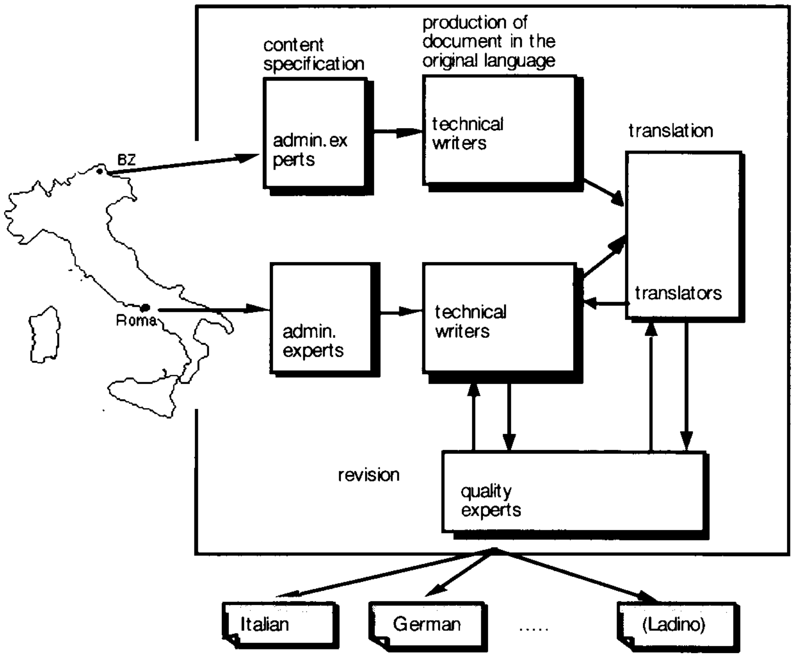


*Figure 1.* Current form production process for the bilingual province of Bolzano.

into German. For documents with national validity, such as laws or forms for pension benefits claim, the original Italian version of texts is first produced in Rome and then sent to Bolzano for translation or, sometimes, revision of the translation. Figure 1 sketches the current information flow in the document production process.

Current procedures for producing equivalent versions of administrative documents in more than one language primarily involve the translation of naturally produced texts. The translated texts are expected to satisfy the same transparency requirements imposed on the source language.

The authors of the original text have primary expertise in a specific administrative domain, whereas the translators have primarily bi-lingual expertise. Often translators work for many different administrative and legal offices and can acquire just a superficial knowledge of technical details. The different specialisations of these two types of document producers often negatively affects the quality of the translations. Translators do not have access to the original technical information and misunderstandings may occur if the initial text they have to translate contains incomplete, incorrect or unclear information. Clarifications with the authors are difficult if not impossible. Moreover, strict time deadlines may affect the refinement and revision phases of the text. This sequence of follow-up steps in the process of document production, with the contribution of various expert groups, introduces long lag-times with consequent delays in the distribution of the final forms.

There may exist additional legal requirements imposed by multilingualism that complicate the translator's work. For example, in the bilingual province of Bolzano, for a certain category of public documents, such as social security service forms, by law the translated text has to appear beside the original text, in the same typographic style, with a paragraph-to-paragraph translation correspondence. This legal parallelism often forces a literal translation of the contents where a different organisation of the text would be preferred in the second language.

## 3. The Automatic Production of Multilingual Administrative Forms

From all the issues related to multilingual form production described in the previous paragraph, it emerges clearly that producing good quality forms is a time-consuming task, requiring good expertise in administrative/legal matters, technical writing for the public and translation. Information technology offers a wide variety of tools that may help and speed up the various form production stages: grammar, style and spelling checkers, translation dictionaries, thesaurus, intelligent editors. Unfortunately all these tools are often restricted to problems at the sentence level and do not support access to the underlying semantic knowledge. Furthermore, they do not help bridge the gap between work done by different classes of experts and cannot solve the problems inherent in it. Also lacking are tools to enable rapid updating of documentation and automatic evaluation tools for assessing the quality of official texts. More advanced technology based on progress in Artificial Intelligence provides a more promising approach to the problem of supporting human writers and translators. Machine Translation and Automatic Generation tools do not simply support the human writer: they themselves propose a possible draft for the texts.

There exist various approaches to Machine Translation (for a comprehensive overview of the field see (Nirenburg, 1987)). The most flexible approach envisages that the translation system receives as input a text in natural language, interprets its content via a lexical, syntactic and semantic analysis and builds an internal representation of the content that abstracts away from the features of the specific language. Starting from this representation level (interlingua), the system chooses the most suitable lexico-syntactic structures conveying the same meaning in another language. The major problem in MT includes understanding the contents of the original text correctly. The automatic system needs to be robust enough to understand all the possible grammatical and lexical structures that appear in the text. Furthermore, since MT tends to be a structure-preserving operation, translated texts often maintain a trace of the source, which may render them less natural and therefore less effective. Recently, semi-automatic translation systems have been developed that allow the user interact with the system to best exploit its potential (in this case, we speak of Machine Aided Translation).

### 3.1. AUTOMATIC GENERATION

An alternative approach to automatic translation is provided by the automatic generation of texts in natural language. A natural language generation system is a computational tool that automatically 'builds' a text (a sequence of sentences) starting from abstract (non-linguistic) specifications. Given the internal representation of the knowledge sources, the system decides what is the relevant information to be communicated, it organises a coherent text structure and produces the most appropriate linguistic expressions to convey the message (Reiter et al., 1995). Automatic generation can be employed profitably to produce technical manuals, descriptions of data automatically collected (e.g., meteorological bulletins), summaries of texts or to produce the system's answers in a human-computer dialogue system (e.g., tutoring systems or intelligent interfaces for data access).

Efforts in the field of natural language generation are more recent in comparison to the more widely investigated areas of machine translation and language understanding. However, in the last twenty years valuable results have been achieved, including the first commercial applications.

In the first generation systems, sentences were assembled by simply putting together predefined portions of text and data from a knowledge base (*canned text*). In recent years, the need for long and coherent texts has focused the research emphasis on higher-level textual and pragmatic concerns related to the generation of multisentential texts, calling for more flexible and powerful architectures. The first system capable of producing coherent multiparagraph texts (TEXT (McKeown, 1985)) was based on the use of *rhetorical schemas*, sort of skeletons of the discourse, indicating in which order the topics must be presented. Schemas are filled in with sentences built according to the information available in the knowledge base. Alternative schemas may be selected to tailor the resulting text to the expertise level of the reader or to repair possible misunderstandings of the reader in previous portions of the discourse. In its simplicity, the approach based solely on rhetorical schemas is particularly effective in producing descriptions of objects from a database, whenever interactive dialogue with the user is not required.

An alternative approach to generating multiparagraph texts is based on the assumption that the coherence of a text can be explained by means of coherence relations (called *rhetorical relations*) which hold between text segments.[1] The Rhetorical Structure Theory (RST (Mann and Thompson, 1987)) proposes a set of meaningful relations, together with conditions that must hold for their presence in a text. In generation, these relations can be exploited to plan the text: a new sentence is added to the plan only if it can be linked to the preceding text by means of a rhetorical relation (Hovy, 1988; Moore and Paris, 1993). Systems that exploit this approach allow the planning of the text sentence by sentence, with the ability

[1] For example, the rhetorical relation of SEQUENCE occurs between two sentences that present events in a temporal sequence, whereas CAUSE links a portion of text describing an event or a state with another portion of text which describes its cause.

to decide on the spot the most relevant information to be included. In this way a dialogue system can more flexibly generate answers taking into account the user's reactions. The most advanced architectures for generation combine the advantages offered by rhetorical schemas and rhetorical relations in an integrated approach, where the global organisation of the text is determined by schemas and the relations control the local coherence of the discourse (Hovy et al., 1992; Not and Pianta, 1995).[2]

### 3.2. MULTILINGUAL GENERATION OR AUTOMATIC TRANSLATION?

In the production process of multilingual administrative documents there exists a moment when the competent office has a complete description of the contents that the new document will have to convey but has not produced the actual text yet. At this stage, the problem is show to choose the most effective automatic tool to help the document production process?

To use automatic translation tools, it is necessary that a first draft of the document is available in one of the target languages, but no formal representation of the contents is required. If the texts are particularly complex, semi-automatic translation tools might be preferred, but in this case human users with competence in translating are required to assist the system computing.

Automatic multilingual generation is a promising approach whenever the content of the documents can be easily formalised in abstract terms or is already available as a result of other elaboration steps (e.g., CAD specifications for a product could be used as input for the generation of the manual describing its functionalities). In this case there is no need for the hand-production of a source text and all the problems related to its interpretation are avoided. The quality of the output documents can benefit from the possibility of organising from scratch the entire discourse and linguistic structure of the text in the most effective way for the different languages. Quality verification and possible post-editing can be done by people with good knowledge of the individual languages, but not necessarily competent in translation methods.

Whenever the content of the documents is too complex for a simple and complete formalization in abstract terms, such as for intricate legal documents, drafting support systems still remain the most promising help tools (Branting and Lester, 1996; Daskalopulu and Sergot, 1995), even though little help is offered in this case to automate multilingualism.

[2] For a more detailed overview of the types of architectures for automatic generation systems see also (Reiter, 1994).

## 4. The GIST System – Generating (Multilingual) Instructional Text

The GIST project (LRE 062-09)[3] addressed the development of a multilingual generation system for the automatic production of texts describing administrative procedures (e.g., the instructions that a citizen has to fill out an application form for pension benefits) in three different languages: English, Italian and German. The GIST system aims at providing good quality drafts of text; such drafts can then be revised and post-edited by professional writers and/or translators. The new scenario, envisaged with the use of GIST's results, avoids the production of a first complete version of the document in one language. The administrative procedure to be communicated is described only once, in abstract terms. Three text drafts are then automatically produced in parallel for Italian, German and English. The considered application domain (instructions for filling out forms) is particularly well suited to the automatic generation task, since – contrary to what happens for less formalizable texts like contracts or laws – usually the structure of forms and their content can be defined in an abstract systematic way (see the following section on the GIST graphical interface).

The advantages introduced in our scenario by the automatic generation of multilingual texts include:

- parallel production of versions of the same document in different languages, avoiding the problems related to translation;
- accessibility of an abstract, non-linguistic representation of the content of the document: this may benefit the work of various expert groups or can be exploited by other connected automatic tools;
- support for the frequent production of similar but not completely identical documents: small changes in the knowledge base or in the abstract message may cause larger changes in the final texts;
- lower document production times, when reusability and small changes of the internal representation are involved;

Another important advantage of using an automatic tool for generating administrative documents is the possibility of encoding directly inside the system the formal requirements on style, syntax, lexicon and text organisation imposed by the expected transparency of the output texts. The user can be allowed to tailor the style of the document by setting appropriate parameters, but it is then a system task to choose the best realisation solutions in a uniform and coherent way. This also guarantees a uniformity in language style across different documents, therefore avoiding the same type of information being encoded in different ways by different

[3] The consortium of the project included academic and industrial partners – IRST (Trento, Italy), ITRI (University of Brighton, England), ÖFAI (Vienna, Austria), Quinary (Milano, Italy), Universidade Complutense de Madrid (Spain) – as well as two user groups collaborating actively to the specification and evaluation of the system – the Italian National Security Service (INPS) and the Autonome Province of Bolzano (PAB). A third user group – the British Department of Social Security (DSS) – agreed to contribute to the project definition and provided valuable input.

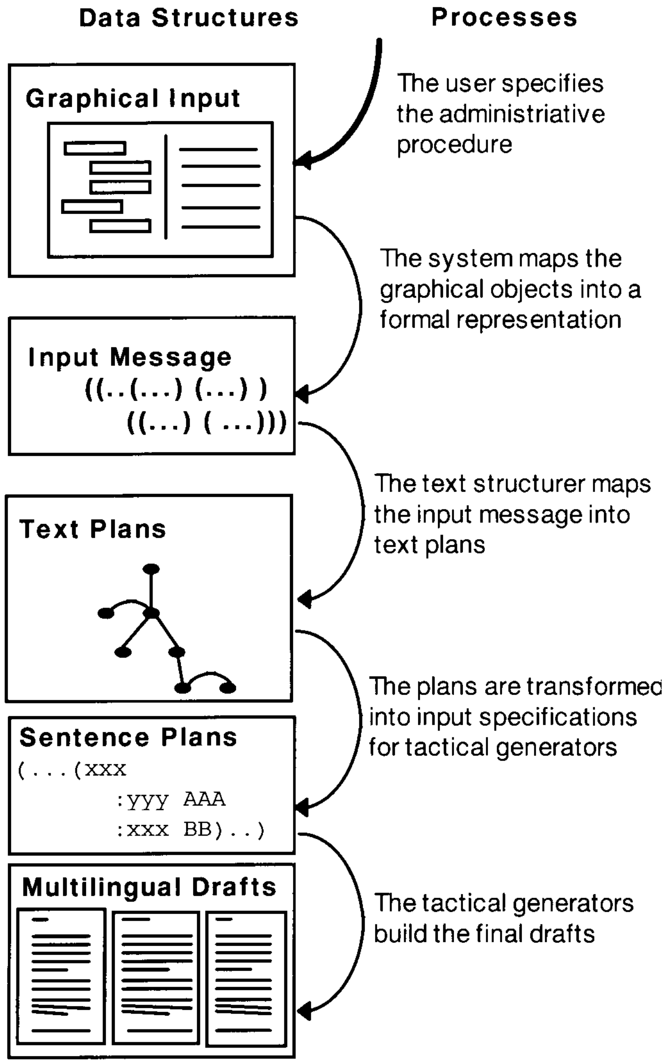


*Figure 2.* A sketch of the generation process.

technical writers, possibly confusing the citizen or slowing down their learning of the form-filling task.

### 4.1. THE GIST FUNCTIONALITIES

Figure 2 sketches how the GIST system produces multilingual texts from the initial message specification up to the final output.

The GIST potential user is the administrative expert in charge for writing a new form. Through a series of menus and buttons, the graphical interface allows the user to specify in a simplified way the structure of the new form: how many fields

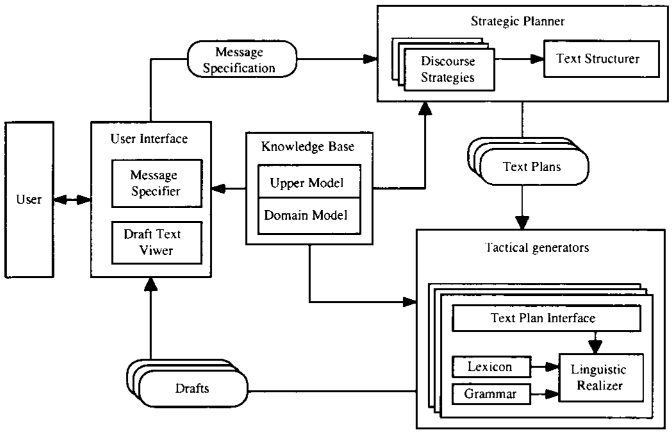


*Figure 3.* Architecture of the GIST system.

should be included in the form, possibly grouping them into sections, and the type of information with which they should be filled. The user can also select some stylistic parameters that will influence the final linguistic presentation, for example the degree of formality in the language used and politeness.

Once the content specification is terminated, the user activates the updating of the system internal knowledge representation: the graphical input is automatically translated into a formal representation (input message) that captures, for each form field, the set of actions to be performed for its correct completion.

This language-independent representation is accessed by the strategic component of the generator to select and organise the relevant information according to the generic text structures and discourse structure relations that are typical for administrative forms. Already at this high level stage of text planning, the system takes into account possible differences in presentation strategies imposed by the different socio-cultural conventions of the three languages considered. The strategic component therefore produces three text plans (one for each language) in the form of trees, whose internal nodes keep the information on the communicative and rhetorical structure of the text, and the leaves represent the semantic content of each sentence, in terms of actions and thematic roles.

The text plans are then translated in a standard formalism (the Extended Sentence Plan Language) and passed to the three tactical generators (one for each language), that select the most appropriate syntactic and lexical structures. The three resulting texts are displayed on the screen. The user can save them and possibly edit them.

### 4.2. THE SYSTEM ARCHITECTURE

An overall sketch of the GIST architecture is shown in Figure 3. Some of the components are built re-using or adapting existing tools, developed by the consortium members inside previous projects or available as public domain software. This is an additional confirmation that the advances in the field of automatic generation have already reached a level of maturity and flexibility allowing their reuse in different applications and domains.

The system consists of six main components: the User Interface, allowing the user to input the content of the message and some global parameters constraining the text generation; the Strategic Planner, building text plans for the three languages; three distinct Tactical Generators, responsible for the linguistic realisation of the text plans; the Knowledge Base where domain dependent (Domain Model) and general linguistic (Upper Model) concepts are defined.

All Tactical Generators are adaptations of existing systems, whereas the graphical input interface and the strategical component have been specifically developed for the project needs.

#### 4.2.1. *The Input Graphical Interface*

For the graphical interface development the goal was pursued of designing a tool that allows administrative experts to express their ideas in a quick and intuitive way, with the possibility of easily recording and reading over specifications, also produced by different people. For this reason the interface was designed to offer a feedback on the form structure, in terms of fields and sections to fill in. All the operations of content specification can be done by selecting buttons and menu options (see the first column of the main frame in Figure 4), helping the user to focus on the relevant concepts and to define a coherent and homogeneous form structure and content (as shown in the second and third columns of Figure 4) (Power et al., 1995).

To help the latter specification, a pop-up window is raised (the smaller gray frame in Figure 4) that allows the user to specify additional conditions that must hold for the correct filling in of a piece of information (e.g., ‘reader is female’).

Users were involved from the first stages of the interface design process. First (paper) mock-up prototypes were shown to experts in document design to identify the most suitable approach for the input specification and other desirable functionalities for the interface. Successive implemented prototypes were evaluated by the same experts: their general assessment was favourable, and suggestions emerged to improve the following versions of the tool.

#### 4.2.2. *Knowledge Representation*

The Knowledge Base is implemented using the LOOM representation language (MacGregor, 1991). To generate the final text in English, German and Italian the GIST system accesses this sole representation of the message and of the domain knowledge, which is independent from the specific target language and provides the adequate abstraction level. Figure 5 sketches how the graphical input in Figure 4 is automatically translated by the system into the internal knowledge representa-

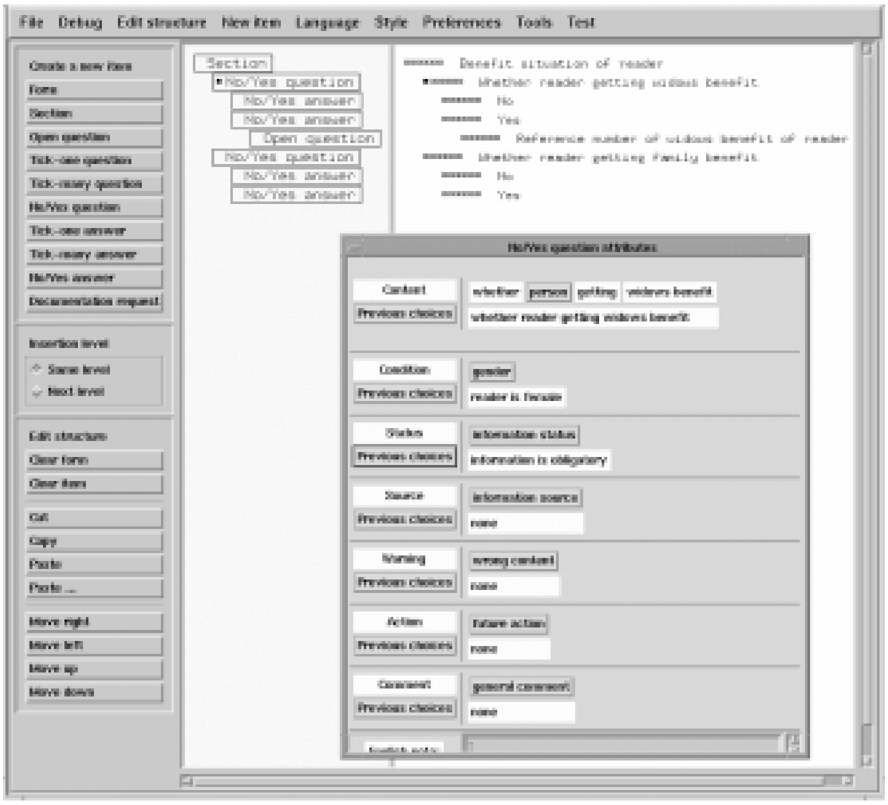


*Figure 4.* Sample input specification. (Screen dump of the GIST graphical interface).

tion. The user is not requested to know the details of the LOOM internal syntax since the graphic interface filters out the low level details of knowledge representation manipulation via menu operations. The knowledge base is structured as a hierarchy of concepts and relations relevant for the domain. The higher levels of the hierarchy (called Generalized Upper Model) allow the classification of the different domain concepts taking into account their potential realisation in the three languages. The Generalized Upper Model (Bateman et al., 1995) has been built from existing Upper Models for English/German and Italian (partly developed within the project).

### 4.3. THE STRATEGIC PLANNER

The central module of the GIST system, the strategical planner (called Text Structurer (Not and Pianta, 1995), was defined using an approach based on the clear distinction of the various knowledge sources coming into play in the planning of a text:

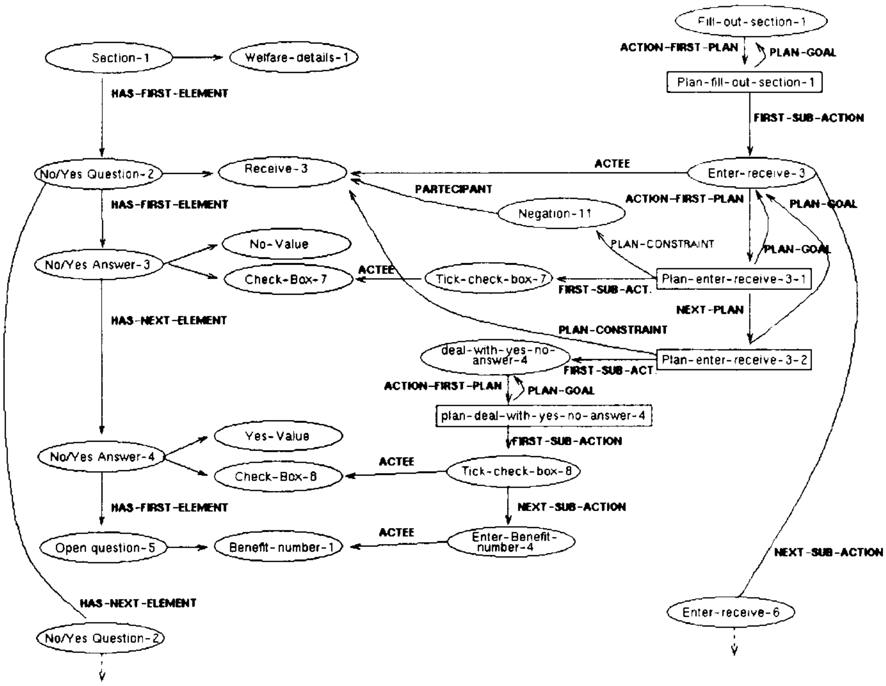


*Figure 5.* Internal representation for the input specified in Figure 4.

- *domain-dependent communication knowledge*, suggesting in which way the content of a text is typically selected and organised in a specific domain (*communicative structure* of the text);
- *general communication knowledge*, suggesting general rules (valid independently of the specific topic of a text) that tell in which way the most appropriate rhetorical relations can be chosen to present the content in a coherent manner (*rhetorical structure* of the text) and in which way anaphoric expressions and marked constructions can be selected to improve the fluency of the text (*cohesive structure* of the text).

The existence of different structural layers (communicative, rhetorical and cohesive structure of the text) was exploited to incrementally build the text plan through subsequent steps integrating new information in the text structure. This three-layer distinction proved particularly suitable for a multilingual setting, since – given a message to be expressed – each structural level can differ in the various languages (Not and Pianta, 1995). Also, the declarative computational resources which guide the generation process were separated from the algorithms that access and reason on them. Declarativity supports easy extensions and modifications of the code and enhances the reusability of information.

The *communicative structure* of the text is expected to represent:

- what the text says explicitly
- in what order
- why

As the first step of our corpus analysis, we examined the bilingual Italian/German texts and the English ones with respect to the order in which the chunks of content are organised to satisfy the communicative intentions. A comparative study was carried out to verify whether differences in content presentation can be observed across the three languages. The exploitation of the observed presentational patterns by an automatic generation tool is important since it reduces for the reader the effort needed to understand the text and increases its relevance. In GIST this information is represented by means of the formalism of *communication schemata* (Not and Pianta, 1995) partially inspired by (Moore and Paris, 1993). A communication schema is a variant of a rhetorical schema (McKeown, 1985) integrated with the intentions that the presentational pattern is meant to satisfy. A communication schema has the following structure:

HEAD: a communication schema descriptor of the type *predicate*($arg_1, \ldots arg_n$) synthetically identifying a plan to achieve a set of intentions.
INTENTIONS: a gloss specifying informally the set of intentions that motivate the use of the schema.
EFFECT: a structure describing formally the main intention of the schema (intention to affect the mental state of the hearer).
BODY: a set of sub-schemata identified by communicative schema descriptors, possibly in the scope of the optional or terminal operators.
CONSTRAINTS: KB-related constraints on the applicability of the schema.
ORDER: constraints on the order of presentation of the sub-schemata introduced in the body.

An example of communication schema is the following:

HEAD: procedure(Action, Type, Actor)
INTENTIONS: 'to get from the applicant some information and possibly to make him/her do some related action'
EFFECT: goal(Actor, give-info(Actor, Action))
BODY: optional(context(Action,Actor)) instructions (Actor, Type, First Instruction)
CONSTRAINTS: first_instruction (Type, Action, First Instruction)
ORDER: before (context, instructions)

The `body` slot of this schema captures the fact that the sequence of instructions that the applicant is expected to carry out may (but need not) be preceded by

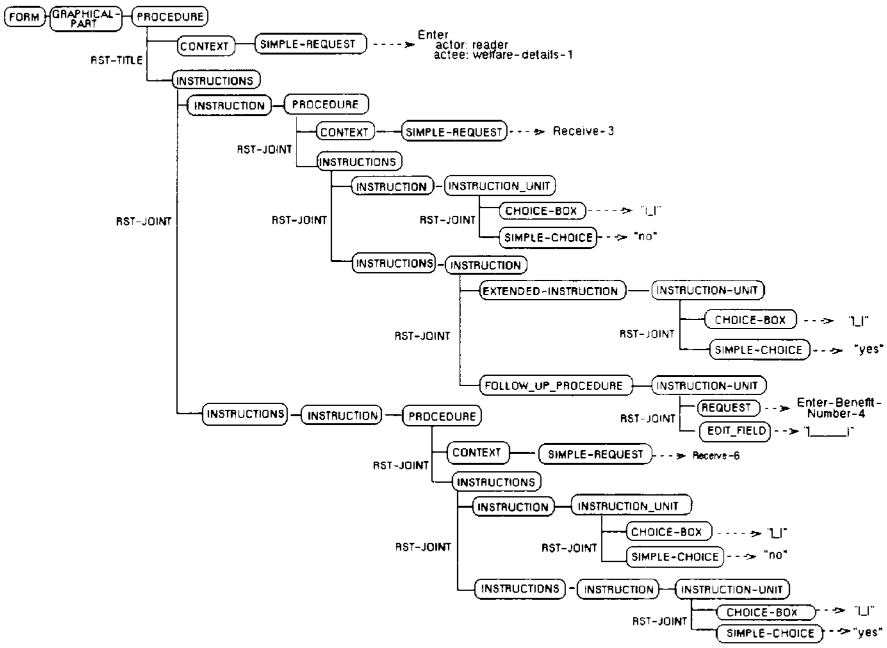


*Figure 6.* Rhetorical plan computed from the knowledge representation in Figure 5.

information (the context) that helps in the comprehension of the instructions (e.g., a definition for a complex term, an explicit motivation for performing the actions). It also states that the context, when present, should precede the instructions (`order` slot). The `constraints` pick out the first sub-action that will realise the procedure.

We assume that each text in our domain can be described by a communicative plan that is produced by the application of one or more nested communication schemata. When a text has to be produced with the communicative goal of letting the reader fill out a form, the text structurer chooses a schema in its library that matches this goal and recursively expands the heads in its body. The choice of the suitable schemata also depends on the information available in the KB.

The obtained communicative structure is then decorated with information on the rhetorical organisation of the output text, such as:

- the explicit discourse relations linking coherently the different text spans;
- the communicative force, i.e. the speech act, for each elementary proposition;
- some phrasing solutions specifying how certain domain concepts tend to be expressed in a certain language.

For the analysis of the discourse relations we took as a reference the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). The choice of this framework is motivated by the fact that it is a well known theory, which has been fruitfully adopted in a number of generation systems, and for which several examples of text analysis are available in literature. Figure 6 shows a communicative structure decorated with rhetorical relations corresponding to the knowledge representation in Figure 5.

Explicit cohesive links are added to the rhetorical structure to enhance the cohesion of the final text. The following factors were taken into account:

- the accurate choice of markers that explicitly signal the presence of a rhetorical relation between chunks of text. For each rhetorical relation the most common realization markers and/or grammatical structures were identified together with the discourse context that supports their choice.
- the use of anaphoric expressions. The most typical linguistic expressions (e.g., pronouns or definite noun phrases) used to refer to the domain entities were identified together with the discourse factors that influence their choice (attentional structure, rhetorical and thematic structure) (Not, 1996).
- an effective thematic progression that guides the reader's attention through the text (Lavid, 1995).

The final text plans (one for each language) are translated in a formal language, the ESPL formalism (Vander Linden et al., 1994), allowing expression of the semantic and pragmatic information required by the tactical generators to produce the appropriate sentence forms. The following sample ESPL formula corresponds for example to the sentence 'Do you receive Widows Benefit?':

```
(ESPLVAR::RECEIVE-3 / DOMAIN-MODEL::RECEIVE
         :TIME PLANNER::E=R=S
         :POLARITY PLANNER::POSITIVE
         :THEME ESPLVAR::READER
         :CLAUSE-TYPE PLANNER::YES-NO-QUESTION
         :NOMINALISATION PLANNER::PROPOSITIONAL
         :POLITENESS PLANNER::NORMAL
         :DISTANCE PLANNER::NONE
         :CONTINUITY PLANNER::DISCRETE
         :ACTEE
             (ESPLVAR::WIDOWS-BENEFIT-2 /
              DOMAIN-MODEL::WIDOWS-BENEFIT
            :DETERMINATION PLANNER::ZERO
            :DEIXIS PLANNER::NONE
            :EMPHASIS PLANNER::NON-EMPHASIZED
            :NUMBER PLANNER::SINGULAR
            :RECOVERABILITY PLANNER::NONRECOVERABLE
            :IDENTIFIABILITY PLANNER::IDENTIFIABLE)
         :ACTOR
             (ESPLVAR::READER / DOMAIN-MODEL::HEARER
            :RECOVERABILITY PLANNER::RECOVERABLE
            :EMPHASIS PLANNER::NON-EMPHASIZED
```

```
        :POLITENESS PLANNER::HIGH
        :DEIXIS PLANNER::NONE
        :NUMBER PLANNER::SINGULAR
        :IDENTIFIABILITY PLANNER::IDENTIFIABLE))
```

#### 4.3.1. *The Three Tactical Generators*

For English, the KPML system, developed at IPSI-GMD as an extension of PENMAN (Penman, 1989), is used as tactical generator. The output produced by KPML for the sample input in Figure 4 is:

```
Welfare details
Do you receive Widows Benefit?
        |_|No
        |_|Yes
               Enter the benefit number.  |_______________|
Do you receive Family Benefit?
        |_|No
        |_|Yes
```

The generator for Italian was developed at IRST and is based on a GB-style unification grammar (Pianesi, 1993). The generator for German, instead, was developed at ÖFAI and is based on a HPSG grammar implemented in the FUF formalism (Matiasek and Trost, 1996). To allow the strategic component to provide a suitable input for modules developed with such different formalisms, a common language to describe the text plans (ESPL) was developed, which is an extended version of the Sentence Plan Language, already successfully used in a number of generation systems.

Figure 7 shows sample output texts produced by the three tactical generators at the end of the overall generation process for an input specification longer than the one provided in Figure 4.

These texts exemplify some of the advantages of multilingual generation when compared to machine translation. Here for example, the system was required to produce the three texts according to the typical style conventions of the administrative language in the various countries. The system built the three texts independently, choosing for each language the most appropriate forms. In the English text, the citizen is more informally and politely addressed with ‘you’ and information requests are more directly expressed with imperative forms. In the Italian and German texts, instead, citizens are more indirectly addressed with the generic term ‘the applicant’ and impersonal forms are used to express requests. In Figure 7 we can also note how auxiliary information associated to the various fields, as in the field ‘date of birth’, in English is put close to corresponding field, whereas in Italian and German it is put in a separate note at the end of the form.

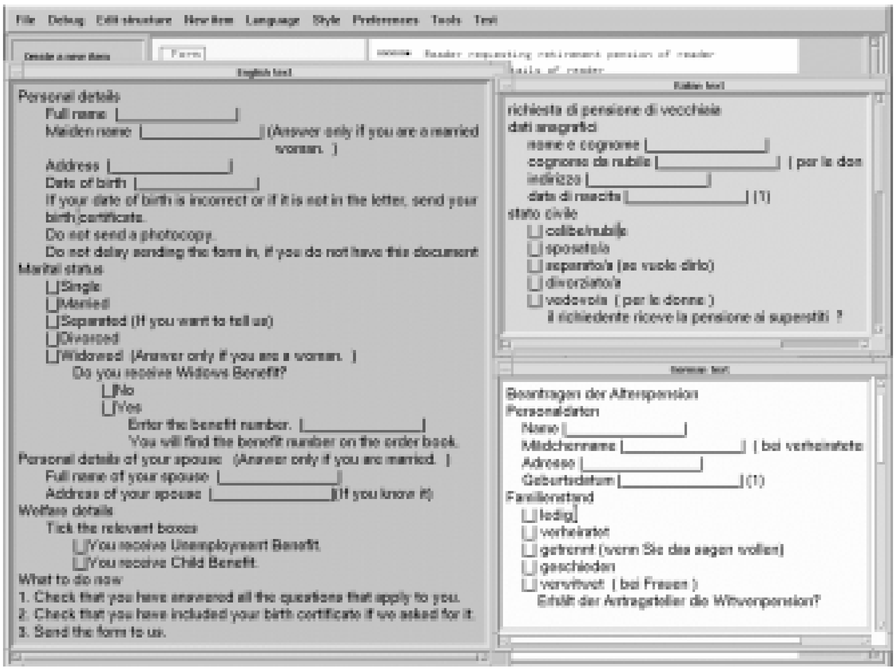


*Figure 7.* Simple examples of output texts.

### 4.3.2. *User Involvement*

To enhance the quality of GIST results, the project aimed at a high user involvement in all the phases of its development. The expertise of administrative experts, professional authors and translators was exploited to identify the features of the application domain and of the target texts. Users were involved, through interviews and experiments, in:

- the choice of the applied domain and the selection of corpus of texts to be analysed;
- the analysis of the system requirements;
- the definition of the quality criteria for the generated texts;
- the evaluation of the intermediate and final prototypes of the project.

An evaluation of the integrated system, including the text drafter and the generated drafts, took place at the end of the project. A group of administrative experts was invited to test the prototype and to comment on nine texts automatically generated. User evaluation results confirmed the usability of the system and therefore confirmed the validity of multilingual generation as an alternative approach to machine translation. Additional effort would of course be requested to put the system into actual practice on the desk of administrative experts, since this requires a substantial extension of the linguistic resources available (e.g., lexicons and grammars) to increase the system coverage.

## 5. Interactive Access to Public Administration Information Systems

Let us now turn our attention to the issue of interactive access to information for the citizen, and introduce the second project.

In a recent overview (Androutsopoulos et al., 1995) of the field of Natural Language Interfaces to data bases (NLI) it was shown that although some of the numerous NLI's developed in the mid-eighties demonstrated positive characteristics in certain application areas, they did not gain the expected wide commercial acceptance. The development of graphical and form-based interfaces, and some intrinsic problems of NLI (e.g., opacity on linguistic coverage and conceptual/linguistic failures) are probably the main reasons for the limited success of commercial NLI to databases. However, the technical advancements in various specific areas of NLP have been notable, and many have resulted in a technology mature for new applications.

The TAMIC[4] (Transparent Access to Multiple Information for the Citizen) project, aimed at proposing a natural language centered interface to substantially improve the way an average citizen can access existing information held by Public Administrations (P.A.). The project was developed in close collaboration with the users of the system, particularly with INPS, the main Italian public entity for social insurance, and the Autonomous Province of Trento. The first phase of the project has achieved two important results: first, research on user needs showed that the adoption of a NLI would be suitable for the situation; second, a demonstrative prototype has been realized and a first consensus for the proposed solution was achieved with its presentation to users and to a good number of P.A. managers.

In the present situation of Italian P.A. offices, a citizen who finds himself needing to access data distributed over several archives must proceed to several successive queries, each of which provides a subset of the needed data, then personally execute the connection of the various gathered data, and finally proceed to distil them and formulate a response for himself. This problem was identified by a variety of studies carried out by the Public Administration in many countries. For example, according to one of these studies, each Italian citizen spends about 20 days per year for bureaucratic matters. Other studies indicate that 25% of the Italian citizens need about 30 minutes to obtain one single piece of information from a public administration office, without taking into account travel inside the city and the fact that often a bureaucratic case needs access to different offices. The creation of integrated information desks will in principle certainly reduce this waste of time.

Yet information systems in Public Administration are characterised by major differences between one another. Navigation in a distributed environment with a

[4] The MLAP-TAMIC EU Project (No. 63-457) consortium included: QUINARY SpA (Prime Contractor), Istituto Nazionale per la Previdenza Sociale (INPS), Istituto per la Ricerca Scientifica e Tecnologica (IRST), Cap Gemini Innovation (CGInn), Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Istituto di Linguistica Computazionale. A follow-up project, named TAMIC-P, was approved by the EU and will start in January 1997.

multiplicity of different information situations is complex even for an operator inside the same administration: to be put in condition to access even a limited number of resources, a worker requires a rather long and costly training period. In creating structures that are not only infra-office but also inter-office, it is easy to foresee an even greater difficulty, one not to be overcome without sophisticated technical support.

TAMIC proposes an attended desk: interaction is mediated by a clerk who interacts with multiple information systems, through the same interface. The clerk can use the interface without knowing the details of the structure of the many different data bases involved. Additionally, clerks, who are usually competent only in some areas of P.A., will be able to provide information also on other topics. The approach stresses that often one single question will not solve the information problem: one needs to install a kind of dialogue to get to the needed information; and it is important that also the citizen understands the dialogue between clerk and computer (that he can monitor on an echo display), so that he can collaborate, for instance providing missing data or exploring alternatives.

With this scenario the goal of a first phase of the project (a second phase is more oriented toward a concrete realization) was to show the developmental possibility of an extended NLI interface that supports:

– data access;
– data interpretation;
– integration of the information with data coming from different data bases.

The prospect is that the interface is such that the interaction allows the clerk to concentrate on the citizen's problem, and that both clerk and citizen can understand and participate in the ongoing dialogue.

### 5.1. USERS ANALYSIS

The first phase of the project was devoted to evaluate how clerks really work in their everyday environment, with the aim of identifying their needs and setting a first group of guidelines for the definition of the TAMIC interface. The study was conducted at the premises of INPS and of the Trento Work Agency; two services of these agencies were investigated: INPS Operative Centers and Trento Work Agency's Professional Orientation and Employment Assistance Center.

The methodology used includes both subjective and objective evaluation of the current situation. Subjective evaluations are carried out interviewing the users of today's systems; objective evaluations derive from observing clerks at work in their everyday environment (ethological observations).

Each technique allows us to detect different relevant information: interviews give a list of problems while observations provide some measures.

Although the study was too limited in the number of people involved to be considered general, investigations have yielded a sketch of the clerks' environment and needs. The study revealed a serious difficulty in the use of present tools, as well as an underutilization of their capabilities.

## 6. Tamic Functionalities

The TAMIC system consists of an environment for the clerk through which he will be able to interact with the underlying information system. It presents itself as a complete interface including a combination of natural language, menus and graphical tools. The final setting for validation in the field will consist of a personal computer-based interface connected to the pension system network and with an echo monitor where the citizen will be able to observe and understand the ongoing dialogue between clerk and system. The citizen will ask questions and interact with the clerk in a clearer way referring to an understandable ongoing dialogue and output data displayed on the screen.

The interface is an integrated environment at the following two levels: (a) at the level of user actions, it permits shifting easily and consistently from one tool to the other; (b) at a deep level it consents maintaining the global focus of interaction. This permits integrating browsing and natural language access into a single global dialogue modality. Along this line TAMIC helps to overcome the problem of the 'teletype approach', provided by earlier NLI that did not exploit the potential of a computer display. These aspects build on experience at IRST (Stock at al., 1995), in the integration of NL modalities (mediated exploration) with graphical browsing of the underlying knowledge (navigational exploration).

The whole TAMIC interface yields the following functionalities (see Figure 8):

• *NL access to structured data.* It allows input of expressions in a relevant subset of natural language (principally complex noun phrase expressions). Linguistic coverage includes quantification, coordination, ellipsis, anaphoric and temporal expressions.

Robustness is an important feature of the interface, so that typing errors or knowledge coverage errors are recovered automatically in large part. The NL expressions can be combined with menu based action indications. Replies by the system are directly manipulable, and through mouse-based actions one gets more details or other linked information.

• *Browsing of the underlying domains using a graphical interface*. The clerk can explore the knowledge available by means of a graphical browser. The portion of knowledge actually displayed is used by the system as a context for interpreting NL expressions.

• *Access to textual data bases*. written material can be retrieved: technical documents (laws, rules, etc. . . . ), which can help the clerk to investigate a particular issue, and material for the citizen that can help him understand a particular matter.

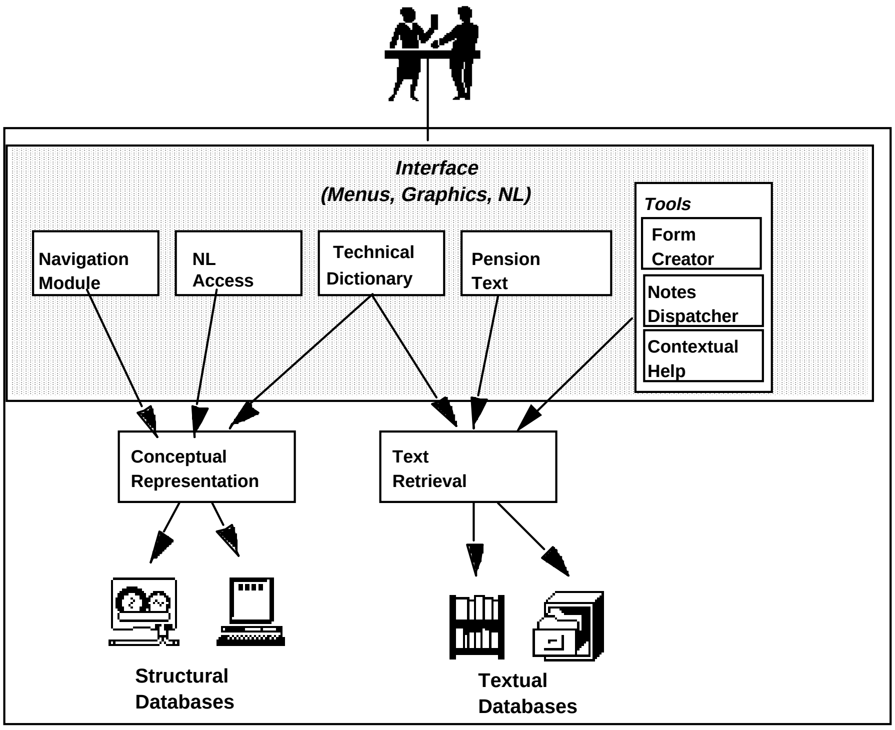


*Figure 8.* TAMIC main modules.

• *Access to a technical dictionary*. A technical glossary including pension terms is accessible by the clerk who needs to know precise term definitions. The same dictionary is also accessible from the navigation interface.

• *Help tools*. TAMIC provides the clerk with a set of tools integrated in the overall interface:

– An on-line contextual help.
– A form creator that allows automatically preparing forms for the individual citizen with relevant data coming from different data bases, so that the form can be printed and given to him for signature, reminder and so on.
– A notes dispatcher to technical offices for cases the clerk is not able to solve.

The interpretation of requests and other utterances as well as navigational interactions results in the interaction with a conceptual model of the domain and of the specific databases. Mapping modules, receiving in input a logical form, produce queries to the relevant data bases and provide the answer material for output presentation.

## 7. The TAMIC Demonstrator

In the preliminary study of the TAMIC project, a demonstrative prototype was developed. The development involved the contribution of some administrations that provide information related to the job market in the province of Trento. In this area, in particular, an office of the Autonomous Province of Trento, the 'Agenzia del Lavoro', operates to bring together job offers and requests. This office manages a large information base distributed over various poles, but needs to integrate its information with that of INPS (the national social security administration), that, for its own institutional needs, manages a large information flow on companies operating locally and in the job sector.

Moreover, the two agencies have similar and complementary aims, as far as job mobility and unemployment are concerned.

The citizen that can access the information point may be a job seeker, a personnel hiring agent, an insured worker, a requester or user of unemployment benefits. The relevant information for the above mentioned sectors was collected considering a number of Public Administration Data Bases. Some of them come from the Autonomous Province of Trento, while others come from INPS.

The demonstrator is based on reuse and adaptation of various modules developed previously at IRST. Only a simulation (i.e., small portions of database information inserted in the knowledge base) of real (i.e., usually non UNIX-hosted and non ethernet/internet connected) databases is realized.

### 7.1. NATURAL LANGUAGE ACCESS MODULE

The natural language access module of the TAMIC system is based on the so called *multilevel semantics* approach, proposed originally in (Scha, 1983) as a powerful architecture for semantic analysis. In this approach, interpreting a natural language sentence is a multi-stage process, starting out with a high-level meaning representation, which reflects the semantic structure of the sentence rather directly. Then translation rules, which specify how the language-oriented semantic primitives relate to those that are used at deeper levels of analysis, are applied. One of the advantages of the multilevel approach is that it allows a natural decomposition of complex tasks and the functional modularization of semantic analysis. In concrete applications the definition of the semantics for a given level has at least two advantages: (i) modules for specific phenomena can be easily introduced within the appropriate level, provided that the module functions contribute to the definition of the semantics for that level; (ii) diagnostics of the semantic analysis is facilitated: in particular, when a sentence is rejected at a certain level, it means that the semantic constraints for that level have been violated (Lavelli et al., 1992).

The overall architecture of the TAMIC NLP system is shown in Figure 9.

The parser used in the TAMIC demonstrator is WEDNESDAY2 (Stock, 1989), a chart-based parser basing its process largely on information stored in the lexicon.

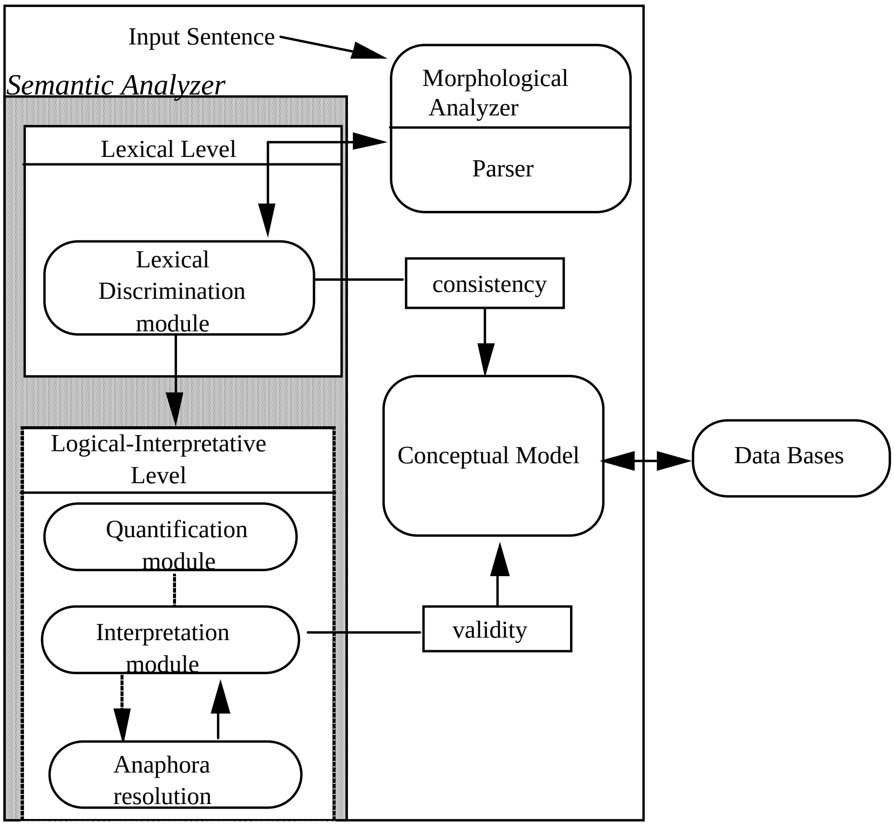


*Figure 9.* The multilevel architecture for the Tamic system.

As far as the semantic analysis is concerned the TAMIC demonstrator includes two levels: the lexical level and the logical-interpretative level. At the lexical level lexical discrimination is computed and a semantic representation of the propositional content of the sentence is produced in output; through lexical discrimination, incrementally only the interpretations satisfying given selectional restrictions are selected. When the propositional content of the sentence is proved to be consistent, the semantic representation produced by this level is passed to the next one; otherwise, if consistency cannot be proved, the whole sentence is rejected.

At the logical-interpretative level specific modules provide a logical form in which both quantifiers and anaphoric ambiguities are resolved. The output is an unambiguous logical form whose predicates are concepts and roles of the Conceptual Model,[5] and whose arguments can be variables and instances.

[5] The structure of the Conceptual Model will be presented in the next section. It is largely defined by means of a knowledge base (KB) using a formalism following the tradition of the kl-one knowledge representation languages (Brachman and Schmolze, 1985).

```
(which x (and (Requirement x)
      (Need-req SW-Engineer x))
    (Have-requirement Mario-Rossi x))
```

*Figure 10.* Logical form produced by the NL access module.

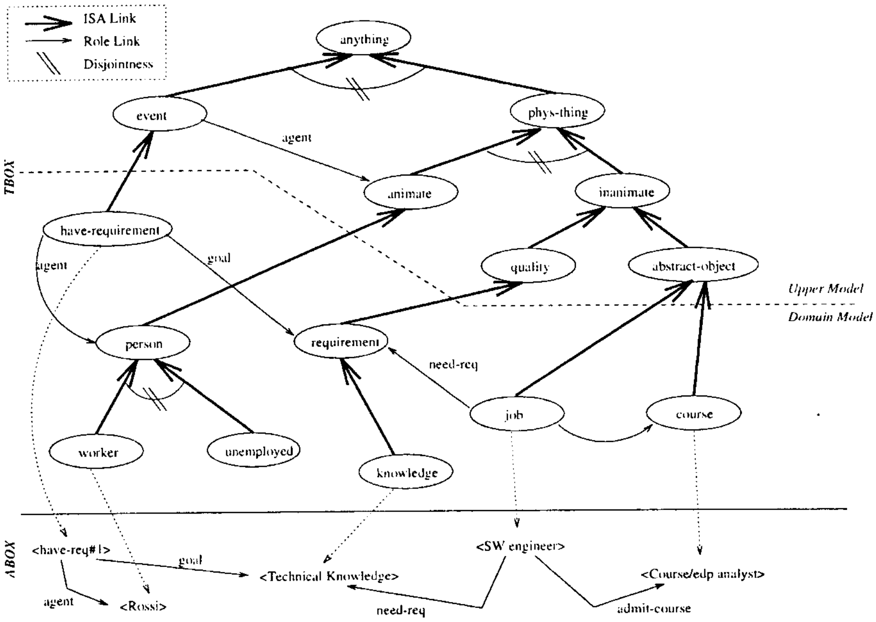


*Figure 11.* A partial fragment of the conceptual model.

As an example, Figure 10 shows the logical form produced by the input sentence 'Which requirements for SW engineer does Mario Rossi have?'. Then this form is translated into a KB query language to retrieve the actual objects in the databases that satisfy the sentence. More details about coupling the Conceptual Model with the underlying databases can be found in (Bagnasco et al., 1991).

### 7.2. CONCEPTUAL MODEL

The sentence meaning is based on a Conceptual Model (see Figure 11), which provides an intermediate representation of the world knowledge, independent from the actual structure of the databases. In previous experiences we have found it useful to separate a general, linguistically oriented conceptual model (the 'Upper Model'), from the conceptual model of the specific application domain (the 'Domain Model'). This separation facilitates the interfacing of the Conceptual Model with the linguistic modules of the system, such as the lexicon and the grammar; moreover, the Upper Model is supposed to subsume specific knowledge of the application domain. We have adopted the ‘Generalized Upper Model’ (GUM), an already existing ontology (see Bateman et al., 1995). The GUM allows incrementally building a representation of the propositional content of the sentence starting from its components by full composition among its concepts and relations. In addition, GUM provides a full set of selectional restrictions, as the basis of the lexical discrimination process. The Domain Model is classified (by means of IS-A links) under the GUM and amounts to about 300 concepts and 300 relations. Both GUM and the Domain Model were implemented using the knowledge representation management systems (KRMS) LOOM (MacGregor, 1991).

### 7.3. MULTIPLE DATABASE ACCESS

The P.A. data access task often involves access to different information sources, both physically distant and structurally not homogeneous. Task complexity is very high, as data contained in different databases are not independent: data pertaining to the same logical subject are actually scattered or replicated into different databases.

Three modules are considered essential for accessing multiple heterogeneous databases from a single NL interface: an integrated Global Data Schema defined on the basis of a series of Local Data Schemata (see Figure 12); a Data Retrieval Engine able to transform a query to the global schema into a set of queries to the Local Data Schemata; an integration of the global data schema and the NL Interface knowledge structure (in TAMIC, the Domain Model), so that a query to the Domain Model is automatically translated into a query to the Global Data Schema.

The definition of an integrated Global Data Schema involves a heavy data modeling phase that must be accomplished in close collaboration with the data administrators. During this phase, the databases are selected and, for each relevant database, a Local Data Schema is acquired or produced; the relationships among the Local Data Schemata are analyzed and formalized and an integrated Global Data Schema is produced.

In TAMIC, these tasks are supported by a suite of graphical software tools, that produce a graphical representation of the Local Data Schema out of a particular data source description (DDL file, flat file, data dictionary, etc. . . . ) and assist the data administrator in building the Global Data Schema and defining its relationships with each Local Data Schema.

Various elements must be taken into account to establish the relationships between each data source and its corresponding Local Data Schema: the kind of data sources (e.g., hierarchical, relational, flat files), the computer environments, and so on. Such information is formalized into ‘Physical Mappings’, information establishing a correspondence between the logical description of a data source and its actual implementation.

The set of transformation applied to the Local Data Schemata in order to obtain a global data schema are in turn encoded as a set of ‘Logical Mappings’. It is

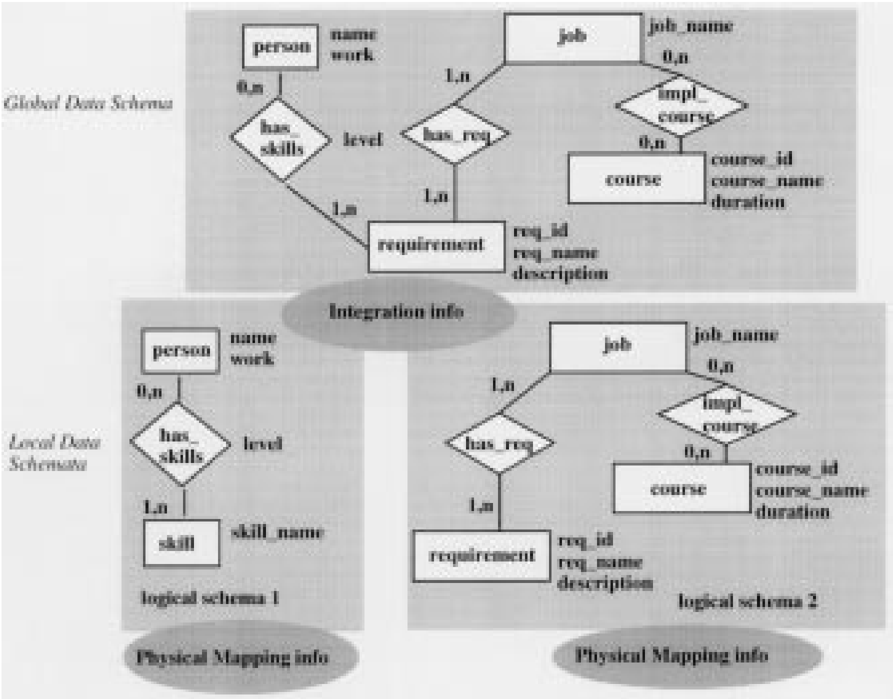


*Figure 12.* Global data schema.

important to underline that the same formalism must be chosen to describe both the local and the global schema. In TAMIC, we have chosen a variant of the Entity-Relationship formalism. Each entity/relationship in the Global Data Schema integrates the information represented by one or more entities/relationships belonging to different Local Data Schemata. Information about the global schema, the physical and the logical mappings are recorded into a relational database, accessible both from the Data Retrieval Engine and from the NL understanding module.

The Data Retrieval Engine uses the Logical Mappings to split a query to the Global Data Schema into a set of queries to the Local Data Schemata. Queries to each Local Data Schema are then compiled into a proper set of query programs and executed in the corresponding computer environment exploiting the Physical Mappings. Both the Logical and Physical Mappings are then used in a backward manner to integrate query answers.

Once given a global representation of the database inter schema, the task of integrating the Domain Model and the Global Data Schema must be afforded, to allow formulating and answering queries via the Knowledge Representation query language. In the TAMIC modular architecture this task involves *coupling* the Domain Model KB with the single virtual relational database. This means that primitive concepts and roles in the KB are made to correspond respectively to unary and binary tables in the global virtual DB (Borgida and Brachman, 1993). In particular, the so called *tight coupling* technique was preferred to the so called *loose coupling* approach (Devanbu, 1993; Borgida and Brachman, 1993), that requires a pre-loading of the data from the DB into the KB, because *tight coupling* implements DB *on demand* access.

### 7.4. DIALOGUE MANAGEMENT

In the TAMIC system, the dialogue manager is meant to fulfill the task of constraining inference in the anaphora resolution process and, in general, to provide a coherence notion for information access dialogues.

Grosz and Sidner (1986) have proposed a discourse model with three components: the linguistic structure, the attentional structure (or global focus) and the intentional structure. In a dialogue system for information access such as TAMIC the attentional status is of critical importance. TAMIC uses the dialogue model presented in (Zancanaro et al., 1997), based mainly on the Grosz–Sidner model and on the centering model (Grosz et al., 1995). Below we introduce briefly the main ideas.

Most of the user utterances in an information access system are questions and commands so we can take as the main element of the interaction the *turn*, i.e., the adjacency pair Question/Answer or Request/Response. Each turn is associated to a set of *centers*, that correspond, roughly, to the KB entities referred by some noun phrases in the sentences of the turn. There are some important relations among centers of two different turns. In particular, in an information access dialogue we can imagine that a question is coherent with the preceding dialogue when it is a follow-up of a previous turn. There are at least two ways in which a question may be a follow-up of a turn: either it is about the same subject of the question of that turn, or it is about the same subject of its answer. Our model exploits also multimodal communication contributions to the dialogue: the multimodal context and every non-linguistic action modify the attentional state maintained by the system.

## 8. An Example of Interaction with TAMIC

As pointed out, the idea proposed is that of a window presided over by an employee who can access a number of data bases (multifunctional window) through a single interface. The employee interacts with the system in Italian, using the keyboard, and can express himself with linguistic flexibility. A dialogue develops (not just isolated sentences) with the computer, without the necessity of the employee knowing the data structures in detail. The citizen can follow the interaction on a second screen.

Here is a session with the system, including the communication between citizen and clerk. An unemployed citizen awaiting the outcome of his subsidy request,


*Figure 13.* The citizen presents himself to the clerk: "I am Giuseppe Rossi. I have requested unemployment compensation and would like to know the status".

comes to the window to check the status of his request and the possibility for his professional requalification in the job market.

The operator asks the citizen to identify himself: treasury department code number or first and last name and possibly date and place of birth. After the citizen confirms the data found by the system, the operator asks TAMIC: '`What type of unemployment request has been made?`'

TAMIC communicates that there are two unemployment requests, of which one was granted and one denied. It opens the menu for the accepted request and it communicates the amount to be given, starting date and length of time support will be given. As for the request denied, it communicates the brief reason found in the menu, which says 'Denied for lack of insurance prerequisites'. The citizens reads the communications and asks: 'Why was the request presented to INPS refused?'

Let us consider two different cases: in the first case we have at the window a clerk with little experience in the employment area; he needs to consult the rules about the specific matter. In the second case the clerk is an expert in this area and can ask directly for the relevant information.

*Non-Expert Operator*

Selects the 'Help' button from the control bar.

TAMIC's aids are context-sensitive: choosing 'Help' from the control bar brings up information about items the operator might need, chosen on the basis of previous interactions. In our case the help system makes some text cards on unemployment available. In particular it is shown that the rules of employment institute establish that the worker must have at least 12 months on the job to be able to use the insurance. The operator decides to check this constraint on the insurance paid. He asks TAMIC: '`Show me the insurance situation`'.

A data sheet appears containing the details of the insurance position of Giuseppe Rossi, subdivided into contribution periods. The clerk checks the contributions for each individual period of work. He explains to the citizen that his last job lasted less than the 12 months requested by the rules on unemployment.

*Expert Operator*

A clerk who is expert in the unemployment sector, and therefore knowledgeable about the matter, asks: '`Calculate the months spent working in the last job`'.

The citizen reads the reasons the request for unemployment benefits was denied by INPS, then he asks: 'What are the job offers at present?'

Before searching the job offers, the clerk finds it useful to check how much time remains before the citizen reaches the age at which he may go on pension. He asks the system: '`How long until he may be pensioned?`'

TAMIC shows the period remaining, expressed in months and years. This is calculated on the basis of previous contributions and prevailing rules.

The clerk tells the citizen the time remaining. To correctly orient the citizen in his search for work he asks TAMIC to list his previous work experience: '`What were his previous work experiences?`'

The clerk asks confirmation from the citizen of the qualifications obtained. Giuseppe Rossi confirms and asks: 'I would prefer to find a job as cook'. The clerk asks TAMIC: '`Which entities requesting a cook meet his requirements?`'

There appears a menu that includes a large number of hotels and restaurants looking for a cook. The name of each entity is a button that once pushed with the mouse opens an informative file. The entities found are numerous, therefore the operator seeks to reduce them by asking the citizen for some preferences: '*Is the location important for you?*' The citizen replies: 'I prefer the Riva del Garda area'.

TAMIC is told: '`Only those in Riva del Garda`'. On the preceding menu only the entities in the Riva del Garda area appear. The entities proposed by TAMIC are still numerous. Probably many requests for cooks are seasonal, therefore the clerk asks the citizen: '*Would you accept seasonal work?*' Giuseppe Rossi replies: 'I prefer not'. The operator asks TAMIC: '`Which positions are not seasonal?`' Only one entity remains, which is shown by TAMIC . 'Good. How may I make a job application?' The operator tells TAMIC: 'Fill out the job application form'. TAMIC may automatically put data into previously established forms.

In this case a personalized job application form with Giuseppe Rossi's data and addressed to the entity selected is produced.

## 9. Tamic Development Status

The description of the TAMIC project presented in Sections 7 and 8 refers to a one-year feasibility study performed in 1995. The implementation of the demonstrator was based on the partial reuse of pre-existing modules; access to the user archives was simulated. Afterwards a two year project, TAMIC-P (Transparent Access to Multiple Information for the Citizen – Pensions, LE-4253), started in January 1997 and will be going on until December 1998. The aim of TAMIC-P is to support a desk operator of a pension organisation in providing citizens with the relevant information about their pensions and social security, by designing a system which allows a transparent and efficient access to multiple databases and textual documentation. The system interface will allow the use of natural language together with standard WIMP techniques both to search for specific data and for exploration of information. All the phases of the development will be verified through tests based on user centered methodology. The final prototype, fully integrated in the user environment and with real access databases, will be available at the end of the two years.

## 10. Conclusions

We have presented two projects concerned with the application of natural language processing technology for improving communication between Public Administration and citizens. The first project, GIST, is concerned with automatic multilingual generation of text containing instructions for form-filling. The second project, TAMIC, aims at providing a tool for interactive access to information, centered on natural language processing and supposed to be directly used by the clerk but with participation in the interaction on the part of the citizen. Both projects would need further development to be experimented in the field. For the moment, GIST has successfully passed its final review as a European sponsored project; we are considering various options for building upon these results. The TAMIC initial work has now led to a more specific project, called TAMIC-P, under the Language Engineering Programme and focused specifically on the Pensions domain. The project has as prime contractor INPS, the Italian pension system, and has been sponsored by the top management of the organization: a clear sign of commitment on the part of a giant P.A. entity.

## Acknowledgements

We would like to thank Fabio Pianesi, coordinator of GIST, the other IRST people that have taken part in the projects described here, Paolo Bresciani (TAMIC), Alberto Lavelli (TAMIC), Daniela Petrelli (TAMIC), Emanuele Pianta (GIST), Fabio Rinaldi (GIST), Massimo Zancanaro (TAMIC) and all partners and users involved in the projects.

## References

Androutsopoulos, I., Ritchie, G.D., and Thanisch, P. (1995). Natural language interfaces to databases – an introduction, *Natural Language Engineering* **1**(1), 29–81.

Bagnasco, C., Bresciani, P., Magnini, B., and Strapparava, C. (1996). Natural language interpretation for public administration database querying in the TAMIC demonstrator, in: R.P. van de Riet et al. (eds.), *Application of Natural Language to Information Systems*. IOS Press, Amsterdam.

Bateman, J., Magnini, B., and Fabris, G. (1995). The generalized upper model knowledge base: organization and use, in: *Proceedings of KB-KS-95, Second International Conference on Building and Sharing of Very Large-Scale Knowledge Bases*, April 9–12 1995, Twente, The Netherlands.

Borgida, A. and Brachman, R.J. (1993). Loading data into description reasoners, in: *Proceedings of ACM SIGMOD*.

Brachman, R.J. and Schmolze, J.G. (1985). An overview of the KL-ONE knowledge representation system, *Cognitive Science* **9**(2): 171–216.

Branting, L. Karl and Lester, James C. (1996). A framework for self-explaining legal documents, in: *Proceedings of the Ninth International Conference on Knowledge-Based Systems* (Jurix-96), Tilburg University, The Netherlands.

Daskalopulu, A. and Sergot, M. (1995). A constraint-driven system for contract assembly, in: *Proceedings of the Fifth International Conference on AI&Law*, ICAIL'95, University of Maryland, ACM Press.

Devanbu, P.T. (1993). Translating description logics to information server queries, in: *Proceedings of Second Conference on Information and Knowledge Management* (CIKM'93).

Grosz, B. and Sidner, C. (1986). Attention, intention and structure of discourse, *Computational Linguistics* **12**(3), 175–204.

Grosz, B., Joshi, A., and Weinstein, S. (1995). Centering: a framework for modelling the local coherence of discourse, *Computational Linguistics* **21**(2), 203–225.

Hovy, E., Lavid, J., Maier, E., Mittal, V., and Paris, C. (1992). Employing knowledge resources in a new text planner architecture, in: R. Dale, E. Hovy, D. Roesner and O. Stock (eds.), *Aspects of Automated Natural Language Generation, Proceedings of 6th International Workshop on Natural Language Processing*, Springer Verlag.

Hovy, E.H. (1988). Planning coherent multisentential text, in: *Proceedings of ACL '88*.

Lavelli, A., Magnini, B., and Strapparava, C. (1992). An approach to multilevel semantics for applied systems, in: *Proceedings of the 3rd Conference on Applied Natural Language Process of the ACL*, Trento, Italy.

Lvid, J. (1995). From interpersonal options to thematic realisation in multilingual administrative forms, in: *Proceedings of the Workshop on Multilingual Generation at IJCAI'95*, Montreal.

MacGregor, R. (1995). Inside the LOOM description classifier, in: *SIGART Bulletin* **2**(3), 88–92.

Mann, W.C. and Thompson, S.A. (1987). Rhetorical structure theory: a theory of text organization, in: L. Polanyi (ed.), *The Structure of Discourse*, Ablex Publishing Corporation.

Matiasek, J. and Trost, H. (1996). An HPSG-based generator for German. An experiment in the reusability of linguistic resources, in: *Proceedings of COLING'96*, Copenhagen, Denmark.

McKeown, K.R. (1985). *Text Generation – Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*, Cambridge University Press.

Moore, J. D. and Paris, Cecile L. (1993). Planning text for advisory dialogues: capturing intentional and rhetorical information, in: *Computational Linguistics* **19**(4), 651–694.

Nierenburg, S. (1987). *Machine Translation. Theoretical and Methodological Issues*, Cambridge, Cambridge University Press.

Not, E. and Pianta, E. (1995). Issues of multilinguality in the automatic generation of administrative instructional texts, in: *Topics in Artifical Intelligence*, Proceedings of the Fourth Congress of the Italian Association for Artifical Intelligence, AI∗IA '95, Firenze, Italy, 11–13 October 1995, Lecture Notes in Artificial Intelligence, Springer.

Not, E. (1996). A computational model for generating referring expressions in a multilingual application domain, in: *Proceedings of COLING'96*, Copenhagen, Denmark.

*The PENMAN User Guide*. (1984). Information Sciences Institute, Marina del Rey, California.

Pianesi, F. (1993). Head-driven bottom-up generation and government and binding: a unified perspective, in: H. Horacek and M. Zock (eds.), *New Concpets in Natural Language Generation*, Pinter Publishers.

Power, R., Cavallotto, N., and Pemberton, L. (1995). *The GIST Specification Tool*, GIST Deliverable PR-3b, LRE project 062-09.

Reiter, E. (1994). Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?, in: *Proceedings of the Seventh International Workshop on Natural Language Generation* (IWNLG'94).

Reiter, E., Mellish, C., and Levine, J. (1995). Automatic generation of technical documentation, in: *Applied Artificial Intelligence* **9**, 259–287.

Scha, R. (1983). *Logic Foundation for Question Answering*, Technical Report MS12.331, Philips Research Laboratory.

Stock, O. (1989). Parsing with flexibility, dynamic strategies and idioms in mind, *Computational Linguistics*, **15**(1), 1–18.

Stock, O. and the AlFresco project team. (1993). AlFresco: enjoying the combination of natural language processing and hypermedia for information exploration, in: M. Maybury (ed.), *Intelligent Multimedia Interface*, AAAI Press/ MIT Press, Menlo Park, CA.

Stock, O., Strapparava, C., and Zancanaro, M. (1995). Explorations in a natural language multimodal information access environment, *Workshop on Intelligent Multimedia Information Retrieval, IJCAI 95*, Montreal.

MLAP TAMIC Project 63057. (1995). Technical Report.

Zancanaro, M., Stock, O., and Strapparava, C. (1997). Multimodal interaction for information access: exploiting cohesion, *Computational Intelligence* , **13**(4), 439–464.