Multilingual Summarization by Integrating Linguistic Resources in the MLIS-MUSI Project

Alessandro Lenci[1], Roberto Bartolini[2], Nicoletta Calzolari[2], Ana Agua[3], Stephan Busemann[3], Emmanuel Cartier[4], Karine Chevreau[4], José Coch[4]

Università di Pisa, Dipartimento di Linguistica[1]
via Santa Maria, 36, 56100 Pisa, Italy
alessandro.lenci@ilc.cnr.it

Istituto di Linguistica Computazionale – CNR[2]
Area della Ricerca, via G. Moruzzi 1, 56100 Pisa, Italy
{roberto.bartolini, glottolo}@ilc.cnr.it

DFKI GmbH[3]
Stuhlsatzenhausweg 3, D-66133 Saarbrücken, Germany
{agua, busemann}@dfki.de

Lingway[4]
33-35, Ledru-Rollin, 94200 Ivry, France
{emmanuel.cartier, karine.chevreau, jose.coch}@lingway.com

Abstract

In this paper we will illustrate the approach to multilingual automatic abstract production adopted by the EU-sponsored project MLIS-MUSI. Although a small scale research project, MUSI has tried to tackle the challenges set by multilingual summarization by adopting an original approach based on the definition of a shared ontology and representation language, and on the reuse of existing linguistic resources. MUSI combines a statistic-based module for relevant sentence extraction and a concept-based component to generate multilingual summaries.

1. Introduction

The huge amount of documents available – either in the Internet or the Intranet – calls for computational tools supporting search and filtering of information. Research in *multilingual summarization* tries to tackle this bottleneck by pointing at two crucial conditions for improving the access to digital content. First of all, users must be able to have a quick access to (at least) the information that is critical for them to decide about the document relevance with respect to their needs. Moreover, it is essential to provide users with this information encoded in their own language, independently of the source language of the document.

On the other hand, automating the production of multilingual abstracts and short summaries is widely recognized as a highly challenging task (Mani and Maybury 1999). The difficulties concern both the definition of a satisfactory algorithm to determine the relevance of text portions, and the process of generating multilingual texts based on information coming from possibly different source languages.

In this paper we will illustrate the approach to multilingual automatic abstract production adopted by the EU-sponsored project MLIS-MUSI (*Multilingual Summarization for the Internet*). Although a small scale research project, MUSI has tried to tackle the above challenges by adopting an original approach to multilingual summarization based on the definition of a shared ontology and representation language, and on the reuse of existing linguistic resources. The latter include computational general-purpose lexicons, shallow parsers (chunkers and dependency analysers) and text generators, which have been interfaced and partially adapted to the new task.

In the next section, we will outline the MUSI system architecture. In section 3 the algorithm adopted to select the relevant text portions will be described. Section 4 will be devoted to describe the internal conceptual representation in MUSI and the resources for robust NLP that have been customised to perform the linguistic analysis of source sentences and the summary generation. In section 5, some preliminary results of the system evaluation will be presented.

2. The MUSI architecture: general overview

In the field of text summarization, there is a widespread consensus on the use of statistics-based techniques (Many and Maybury 1999), which are inherently domain independent and avoid the problems of grammar coverage. In this view, summarization is seen mainly as a process of sentence extraction and concatenation. On the other hand, purely statistical approaches can difficultly adapt the summaries to different user needs and are unable to create cross-lingual summaries.

The language barrier can be crossed by providing a representation of the document content at the conceptual level. This also opens up opportunities for user-oriented summary formulation, and information fusion – essential for multi-document summarization – becomes possible. However, concept-based techniques suffer from other, well-known drawbacks, such as the dependence on domain and linguistic knowledge, and consequently their effective lack of scalability.
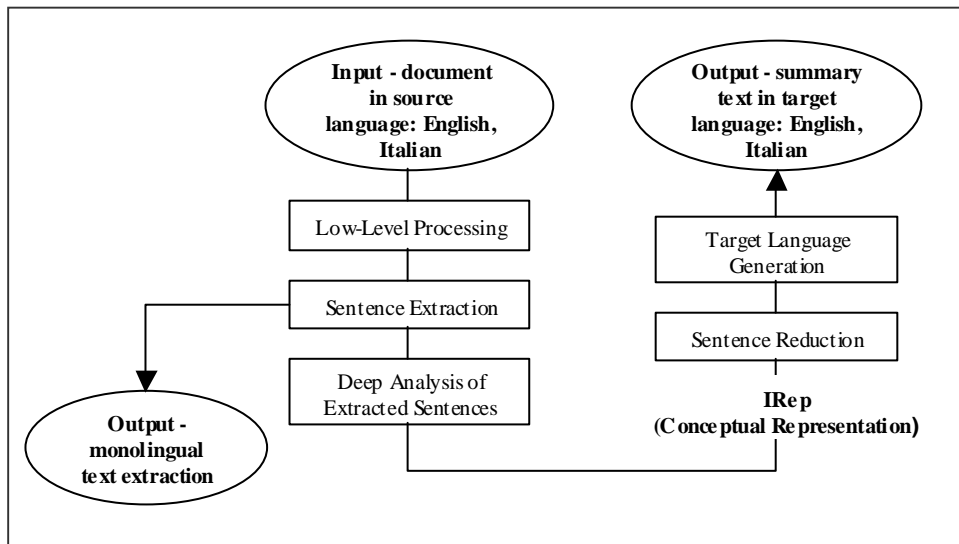
Figure1: MUSI architecture

MUSI approach to multilingual summarization tries to combine both these views in an algorithm that includes i.) *a statistic-based module for relevant sentence extraction* and ii.) *a concept-based component to generate multilingual summaries*. More specifically, MUSI consists of a prototype system for multilingual summarization integrated within a commercial Information Retrieval platform (Lexiguide®), provided by one of the project partners. Summarization is here intended as a "query-biased" process of sentence extraction (see Tombros & Anderson 1998), where users' queries represent one of the main parameters to identify relevant sentences. Currently the MUSI system takes as input HTML-encoded English and Italian scientific articles. The latter are derived from *The Journal of Anaesthesiology*, an English-Italian bilingual medical journal available on-line. The main output is represented by short summaries in German and in French.

MUSI is based on a pipelined architecture of five modules (Figure 1). These are linked up by intermediate XML representations that accumulate the results of the successive processing steps. A first phase of low-level processing of the input document feeds the sentence extraction algorithm. At this stage – as another output of the system - the original documents are offered with the most relevant sentences highlighted. In the next step of deep analysis, robust NLP tools identify the linguistic structure of the selected sentences, necessary to map them onto the conceptual representation level (IRep4). The latter provide the input to the summary generation, after a step of sentence reduction. The generation module produces indicative summaries (in the sense of Sparck-Jones 1999) of the document content. Summaries include both translated portions of the extracted sentences, and "meta-statements" about the original document. The latter provide the user with additional optional information about the content and structure of the source text, the relevance of the extracted pieces of information and of the whole document with respect to the query, etc. Users can customize the summary length, as well as some other aspects concerning style and presentation. A sample French MUSI summary from an English text is reported in Figure 2.

Crossing the language barrier by translating the selected sentences creates a problem of scalability that does not exist with "mono-lingual" sentence extraction techniques. However, this drawback is in many applications outweighed by the additional options for custom-tailoring texts offered by the usage of dedicated generation components. Such options are not available with sentence extraction techniques.

3. Relevant sentence selection

The selection of relevant sentences depends on four parameters: i.) the presence of *cue-phrases*; ii.) how the sentence matches the user-query; iii.) the sentence position within the text; iv) the number of sentences to select.

The sentence extraction process is composed of two modules. The first one converts the English and Italian HTML input file into an XML file with structure tags (titles, subtitles, sections, subsections, paragraphs, sentences). The second module annotates sentences with attributes specifying the presence of cue-phrases, and query instances, and the sentence position. These attributes are used by the *MUSI Summarizer*, which consists of a weighting module and a sentence extractor. Sentences are weighted according to position, query instances and cue-phrases:

1. **cue-phrase** identification is carried out through regular expressions applied on the sequence of sentences and titles of the processed text. Each cue-phrase is associated to a weight assigned to the matched sentence. A number of cue-phrases has been selected for Italian and English, on the base of their ability to identify the sentence relevance. Cue-phrases in MUSI are partly generic and partly domain specific, and have been classified in terms of the rethorical function they express (e.g. Thematic Announcement, Conclusion, Definition, etc.). For example, the following pattern is used to identify English sentences expressing a Thematic Announcement:

Figure2: A sample MUSI summary

```
(present|following|my|our|^(this))(+)(doc
ument|article|paper|review|work|update|pa
rt|section|paragraph)(  +)(aim  at|aimed
at|aims at|is aimed at…)
```

2. The **query instance processing** matches a boolean query (using AND and OR operators) in the sequence of sentences and titles of the processed text and assigns a given weight to the matched sentences.
3. The **sentence position processing** assigns to each sentence a weight relative to its position in the text. For instance, sentences in introduction and conclusion sections are assigned a higher weight, as well as the first sentence of each section.

The sentence extractor module assigns to each sentence a global weight, computed on the basis of the above parameters. Then, the n most weighted sentences are selected, with n being some threshold specified by the user. The relative weight of the cue-phrase, query and position parameters can also be varied by the user, to tune the sentence extractor according to its specific needs.

The selected sentences feed the process of multilingual summary generation. At the same time, they are presented to the user as highlighted text portions in the source HTML file, and form the monolingual output of MUSI.

4. Integrating linguistic resources: text analysis and summary generation

Linguistic analysis applies to the output of the algorithm described in the preceding section, to produce the conceptual representation (IRep4) of the sentence content. The conceptual representations feed the text generators in charge of producing the summaries in the target languages. Both the analysis and the generation steps are performed by applying existing, independently developed robust NLP tools. These include a finite-state, shallow dependency analyser for Italian (IDEAL) a syntactic parser for English (LexiQuest Sentence Parser), a robust dependency-based concept-to-text generator for French (LexiGen, formerly named AlethGen; Coch and Chevreau 2001) and a generator based on production rules (TG/2) used in MUSI with a grammar of German.

The issue of existing resource integration and customisation surely represents an independent added value of the project. The tools employed in MUSI had previously been developed and/or used for tasks as different as query analysis, environmental report generation, lexical acquisition, etc. In fact, the process of customization undertaken to make them adapt to a general complex task such as multilingual summarization has also provided an indirect task-oriented evaluation of these resources.

Besides, the use of robust and shallow NLP tools also opens up the possibility of partial results: less-than-complete analyses of source sentences can nevertheless be processed by the generators. In fact, this is consistent with the idea of producing indicative summaries. Although the generation of full-blown complex sentences is considered as the optimal result, summaries are also well expected to contain "relevant chunks", such as main sentence nuclei and complex terms related to the user's query, complemented with meta-statements. To this purpose, a scale of satisfactory output types has been defined during the project with the help of domain experts.

A crucial ingredient for the resource integration performed in MUSI is given by IRep4, a concept-based internal representation language designed to represent the information extracted from the analysed sentences and to provide the input to the text generation modules. IRep4 has a twofold role within the MUSI approach to multilingual summarization: i.) to provide an abstract "language-independent" representation of the sentence and term contents, and ii.) to provide the crucial interface among the various analysis and generation components, thereby contributing to their effective functional integration within the general architecture.

In the next subsections, we will outline the essential features of IRep4 and we will provide details on the linguistic analysis and summary generation modules. For reasons of space, we will limit ourselves to describe the Italian analysis tools and the German text generator integrated in MUSI. Finally, some critical reflections on the IRep4 will be presented, in the light of the results of the text analysis and generation modules.

4.1. The conceptual level

The keystone for multilingual summary generation in MUSI is provided by IRep4, a hierarchical predicate-argument structure complemented by a rich variety of features and modifiers (Chevreau *et al.* 2000). The IRep4 basic elements are atomic and predicative concepts, forming an ontology shared by the four project languages. In particular, predicative frames are based on the SIMPLE formal specifications (Lenci *et al.* 2000). The following is the IRep4 assigned in MUSI to the Italian sentence *La malnutrizione rappresenta un fattore progostico negativo in corso di broncopneumopatia cronica ostruttiva* 'Malnutrition represents a negative prognostic factor in chronic obtrusive broncopneumonia'.

```
PROP {
     Type = TYPE_PROP;
     Value = P_ARG1_represent_ARG2;
     CAT = V_SEN;
     Time_Rep = [PRESENT, PRES_USUAL];
     Arg1 = ITEM {
          Type = TYPE_ITEM;
          Value = C_malnutrition;
          DET = def;};
     Arg2 = PROP {
          Type = TYPE_PROP;
          Value = P_factor_of_ARG1;
          CAT = NP;
          DET = indef;
          Arg1 = 0;
          Mod1 = [RESTR, C_prognostic];
          Mod2 = [RESTR, P_ARG1_negative];};
     Mod1 = [TDUR, ITEM {
          Type = TYPE_ITEM;
          Value = C_bronchopneumonia;
          Mod1 = [RESTR, C_chronic];
          Mod2 = [RESTR, C_obstructive];
          Mod3 = [PARENTH, C_COPD];}];}
```

IRep4 expressions derive by the recursive composition of `PROP` and `ITEM` elements, respectively used to represent propositions and terms. These are formally feature structures, whose attribute-value elements express information such argument structure, definiteness, number, tense, co-reference, etc. Although IRep4 is in principle a semantic representation language, its expressions also keep trace of the syntactic categorical properties of the source linguistic elements. To this purpose, the attribute `CAT` is used to specify whether for instance a proposition is realized in the source text as a sentence or as a noun phrase. This information has revealed extremely useful to fill the gap between the language–independent representations provided by IRep4 and the specific needs of text generators.

IRep4 is suitable to represent the semantics of very complex sentences, but at the same time leaves room for various degrees of specification. In fact, co-reference resolution, attachment ambiguities and the incorrect identification of arguments and modifiers are notoriously very common sources of problems, that may lead to incomplete output. To cope with these problems, IRep4 has been designed to integrate possibly underspecified or fragmentary representations. This feature greatly contributes to enhance the robustness of the system and can guarantee a better interface with the text analysis component.

4.2. Text analysis

Deep linguistic analysis is performed by independent modules for English and Italian, which share the I/O formats. This solutions has been adopted in order to maximize the reuse of existing tools for these two languages. Notwithstanding the differences in the specific details of linguistic processing, the Italian and the English modules share the general philosophy of regarding linguistic analysis in MUSI as a robust process of identification of structural information, such as shallow syntactic constituents and functional dependencies, which provide the necessary input to map the source extracted sentences onto IRep4.

The deep linguistic processing of extracted sentences of the Italian texts is performed through an "assembly line" whose main components include:
1. tokenisation of the input text
2. morphological analysis (including lemmatisation)
3. shallow syntactic parsing:
4. chunking (including morpho-syntactic disambiguation)
5. dependency analysis
6. mapping of the language-specific syntactic representation onto MUSI internal representation.

The core of the analysis is represented by a shallow syntactic parsing module, which in turn relies on morphological analysis and morpho-syntactic disambiguation (tagging) as its essential prerequisites. The general architecture of the system adheres to the following principles: 1) modular approach to parsing, 2) incremental analysis, 3) underspecified output (whenever required), 4) cautious use of lexical information, generally resorted to in order to refine and/or further specify analyses already produced on the basis of grammatical information. The system is organised into two different modules, one for *text chunking*, a process of non-recursive text segmentation, the other for *dependency analysis*, aimed at reconstructing the full range of functional relations (e.g. subject, object, modifier, complement, etc.) within sentences.

Text chunking is carried out through a battery of finite state automata, called CHUNK-IT (Federici *et al.*, 1998), which takes as input a morpho-syntactically analysed text, tags the input text and segments it into an unstructured sequence of syntactically organized text units called *chunks*. Chunking requires a minimum of lexical knowledge, namely the entry's lemma, part of speech and morpho-syntactic features. Each chunk is a syntactically organized structure (defined in terms of attribute-value pairs), which encodes chunk-specific features as well as the nature and scope of the dependencies holding between the words *within* each chunk.

Inter-chunk dependencies are identified by another component, IDEAL (*Italian DEpendency AnaLyzer*), which takes in input chunked texts. IDEAL includes two main components: (i.) a Core Grammar of Italian; (ii.) a syntactic lexicon of ~26,400 subcategorization frames for nouns, verbs and adjectives derived from the Italian LE-PAROLE syntactic lexicon. The IDEAL Core Grammar is formed by ~100 rules covering the major syntactic phenomena, including: adjectival and adverbial modification; negation; (non-extraposed)

sentence arguments (subject, object, indirect object); causative and modal constructions; predicative constructions; PP complementation and modification; embedded finite and non-finite clauses; control of infinitival subjects; relative clauses (main cases); participial constructions; adjectival and nominal coordination.

The grammar rules are regular expressions (implemented as finite state automata) defined over chunk sequences, augmented with tests on chunk and lexical attributes. The rules are organized into two major modules:

1. *structurally-based rules*;
2. *lexically-based rules*.

A "confidence value" is associated with some of the identified dependency relations, to determine a plausibility ranking among different possible analyses. Consistently with the principle of incremental parsing, these two modules are to be regarded as two independent and successive steps of analysis. First, IDEAL tries to identify as many dependencies as possible without the aid of lexical information; then lexically-based rules intervene either to refine the output of the preceding step (e.g. by changing the ranking of identified dependencies), or to further specify types of relations based on lexicon look-up.

To represent the dependency structure of a sentence, IDEAL adopts a slightly simplified version of the FAME annotation scheme (Lenci *et al.* 1999), where functional relations are head-based and hierarchically organised to make provision for underspecified representations of highly ambiguous functional analyses. This feature is instrumental in allowing IDEAL to tackle cases where there is incomplete lexical information, or where ambiguous functional relations cannot be resolved (e.g. in the case of the argument vs. adjunct distinction). The following is the ouput produced by IDEAL for the input sentence *In questo caso clinico descriviamo l'incidente occorso a due gitanti a caccia di verdura fresca* "In this clinical case, we describe the accident occurred to two travellers looking for fresh vegetables":

```
Modif(CASO[51],QUESTO[51]<Role=demonstr>)
Modif(CASO[51],CLINICO[52]<Role=restr>)
Comp(DESCRIVERE[53],CASO[51]<Role=IN>)
Subj(DESCRIVERE[53],Pro: P1[53])
ObjD(DESCRIVERE[53],INCIDENTE[54]<Def=1>)
Modif(INCIDENTE[54]<Def=1>,OCCORRERE [55])
Comp(OCCORRERE[55],GITANTE[56]<Intro=A>)
Modif(GITANTE[56],DUE[56]<Role=card>)
Comp(GITANTE[56],CACCIA[57]<Intro=A>)
Arg(CACCIA[57],VERDURA[58]<Intro=DI>)
Modif(VERDURA[58],FRESCO[59]<Role=restr>)
```

The output of both stages of parsing, chunking and dependency analysis, is used to build the IRep4. A conceptual lexicon of about 2,000 entries specifies the possible associations between Italian words and the concepts that form the atomic elements of IRep4. Each entry of ICL has the following tripartite structure:

1. *lemma*;
2. *tests* – specifying the morphosyntactic, syntactic and semantic condition that must hold in the Italian

source text for the lemma to be associated to the concept expressed in the *action* part;

3. *actions* – expressing the concept linked to the Italian word. If the concept is predicative, i.e. has an argument structure, the actions also state the possible syntactic and semantic restrictions on the arguments.

At runtime, for each lexicon entry the relevant tests are performed on the parsing output. The result is a list of words-concept pairs, according to the conceptual lexicon specifications, i.e. the words in the input text for which a mapping onto the MUSI concepts exists. Finally, the full structure of the IRep4 is reconstructed on the ground of the syntactic parsing output, mainly the syntactic dependencies identified by IDEAL. Syntactic relations (e.g. subject, direct object, etc.) are projected onto the IRep4 predicate-argument structures, with the help of the selectional information specified in the conceptual lexicon entries.

4.3. Summary generation

The target languages French and German were covered by different project partners, and hence by different systems. French summaries were generated using a dependency-based framework in the spirit of Meaning-Text Theory called LexiGen (formerly AlethGen). IRep4 is a derivative of a representation language used previously in AlethGen, which rendered the adaptation and reuse of the system comparatively straightforward. In the following discussion we concentrate on the generation of German summaries which required a more profound adaptation of existing resources.

The system TG/2 (Busemann 1996), which was reused for the generation of German, is based on restricted production system techniques that preserve the modularity of processing (the interpreter) and linguistic knowledge (the grammar), which is a major prerequisite for reusing NLG resources across various generation tasks and domains. TG/2 has been shown to be particularly well-suited to generating dialogue contributions and to report generation (Busemann and Horacek 1998). In either case, the input to the system was a non-linguistic representation of domain-semantic facts and actions. For MUSI, the generation of meta-statements was very much in line with these previous usages of TG/2, but the use of fine-grained representations such as IRep4 would expand the application space of TG/2 considerably.

With help of grammars consisting of condition-action rules, TG/2 maps pieces of input structures onto surface strings. If the input languages contain task and domain concepts rather than linguistic semantic expressions, TG/2 does more than traditional realization components: it covers, in a shallow way, all linguistic aspects of the NLG process. Due to the diversity of input languages and the domain-specific, though modest, linguistic coverage required, most of the linguistic knowledge was built from scratch each time. This required less effort than adapting existing broad-coverage resources such as KPML or SURGE would have (cf. Busemann and Horacek 1998).

But TG/2 can implement grammars for very fine-grained input languages, too. This flexibility is achieved by integrating canned text, templates, and context-free

rules into a single formalism. The generation process uses the rules to create a derivation tree, the leaves of which are used to create the system output. A coarse-grained model relies on many canned text parts; a fine-grained model uses many rules to model a particular phenomenon.

Generating a grammar reflecting fine-grained semantics from scratch for a new application is not feasible. Rather such a grammar should be re-used by adapting it to different input languages. In the sequel we briefly describe the challenges encountered for adapting and extending our grammar for a fine-grained semantic model to the needs of MUSI.

4.3.1. Reusing a grammar in TG/2

At first sight, condition-action rules that map input onto linguistic strings have a straightforward interface to the input language: the conditions and the functions accessing pieces of input. Any remaining parts of the rules, such as context-free backbone and feature distribution mechanisms (cf. Busemann 1996) can be kept independent of the input language.

Consider the following rule, which is paraphrased in English and slightly simplified: "If the current category represents the infinite verb complex and the current piece of input does not contain an auxiliary verb stem, requires passive voice and perfect or past perfect tense, then generate the past participle of the verb stem, followed by the string 'worden'". This rule generates e.g. *verwechselt worden* "been mistaken", as in "Mandrake has been mistaken for another plant". Obviously, input-specific ways of testing the conditions of accessing the verb stem are needed. The rules use predicates and functions for this. While the calls can remain invariant, their definitions have to change if a new input language is used.

On closer inspection, this method works only if the nature of tests and access functions is general enough that their values can be computed with different input languages. In the above example, the input language must provide information on tense and voice - e.g. as feature values - or allowing for their cost-effective computation, and it must contain lexical entries. An input language complying to these requirements certainly is highly language-dependent, which is exactly what IRep4 as an interlingua should not be. Hence, IRep4 does not fulfil the requirements. What is more, there is also no straightforward way of deriving them one by one through tests, as they may mutually depend on each other. It turns out that a sentence planning stage had to be introduced that would produce the required information. The sentence planner would transform language-independent semantic IRep4 expressions into language-dependent syntactic structures that in turn can be successfully interfaced to the grammar.

Before we briefly describe the tasks of the sentence planner, let us add a few facts about the grammar. In its final stage it comprised 957 rules covering the main types of sentential structures, as specified by the Duden grammar. The NP syntax comprises prenominal APs (on the basis of adjective subcategorisation frames), generic possessive constructions, a temporal, a locative and an adverbial modifier and a relative clause. In addition, nouns and adjectives can subcategorise for specific arguments. About half of the grammar was newly developed to fulfill MUSI requirements on a general linguistic level. There were 328 new entries to the pre-existing wide-coverage morphological lexicon, most of them technical vocabulary or compound nouns.

4.3.2. Sentence planning

Sentence planning in MUSI consists of lexicalisation and the selection of sentential structures. Usually, the head concept in an IRep4 expression corresponds to the main verb of the sentence. Its lexicon entry determines the possible syntactic realizations of its arguments and, to some extent, its modifiers. There are variations of sentence structure depending on which arguments are actually specified in the IRep4 expression. For instance, if the constituent corresponding to the direct object is missing, a passive version is mandatory. Note that this could also have been found out by virtue of the conditions in the grammar, but it would require extensive testing on properties of constituents. By specifying voice in the output of the sentence planner, the interface to the grammar is kept simpler. Certain modifiers are realized independently of the sentence structure. Their realizations just have to fit in grammatically.

As a general rule, the type of the modifier together with the lexicalisation of its head concept determine the syntactic structure corresponding to the modifier. As an example let us consider RESTR. In Figure 3, it is realized in three different ways. The head concept is a noun, 'case'. The first restrictive modifier is realized as an adjective, which takes prenominal position, whereas the second is a noun, taking postnominal position. In general, nouns give rise to generalized possessive constructions expressed in German as genitive NPs or as PPs with preposition 'von'. The third occurrence RESTR corresponds to a relative clause, as one of its constituents co-refers to the head concept (Coref). Without this specification, an attempt to generate another prenominal constituent might result in 'der unten beschriebene klinische Fall' [the below described clinical case], which is slightly preferable. However, if the content of the modifier is large, such construction tend to become unwieldy and are considered bad style.

```
ITEM{ Coref = J;
      Value = C_case;
      DET = def;
      Mod1 = [RESTR, C_clinical];
      Mod2 = [RESTR, C_poisoning];
      Mod3 = [RESTR, PROP{
              Value = P_ARG1_describe_ARG2;
              Time_Rep = [PRESENT, PERF];
              Arg1 = 0;
              Arg2 = ITEM{ Coref = J; };
              Mod1 = [LOC, C_below]; }]; }
```

Figure 3: The RESTR modifier in Irep4

The phrasal types associated with the sentence structure constrain the lexical choice for their heads. A predicative concept can be realized as a verb, a noun or an adjective depending on its syntactic context. This knowledge is encoded in the conceptual lexicon that was created from scratch for MUSI. It associates every

concept with the possible words, parts of speech, and sentence or phrasal plans. In addition, some semantic properties, such as count/mass noun, and syntactic properties, such as gender, are represented.

The different options are explored in a heuristics-based order until one succeeds. While, in principle, the possibility of a complete failure is not excluded, the set of options was rich enough to produce grammatically and stylistically acceptable output in 95\% of the test cases. The heuristics include heaviness constraints that may lead to clausal instead of phrasal realization, or to extraposition of 'heavy' material.

The resulting syntax-oriented representation is subject to grammar tests. Listing some of the information it represents gives an idea of its nature: the name of the sentence structure, the structures corresponding to clauses and phrases, their nesting, the word stems, and lexical and syntactic properties, such as gender, number, or determiner type.

4.4. Some critical reflections on IRep4

After fully implementing the MUSI sentence analysis and summary generation components, tests showed some systematic flaws and difficulties that could be traced back to compromises in the design of the input language. Let us step back and take a critical look at IRep4.

The atomic elements of IRep4 form a basic *ontology* of concepts which has proved particularly suitable to establish mappings among various language specific terms. This is particularly true if we consider the particular domain of application, in which the rich and complex medical terminology surely represents an obstacle at the multilingual level. In MUSI, each partner has been able to develop its own lexicon fairly autonomously, once the common ontology of concepts have been selected, and new languages could be similarly plugged-in. Moreover, the use of a shared ontology for crosslingual summarization can also benefit from existing available terminological conceptual repositories, e.g. UML for the medical domain.

However, IRep4 is not a mere ontology of concepts but also a full-blown interlingua. On the analysis side, problems have emerged on how to map various types of PPs onto the IRep4 semantic modifier types. In fact, in many cases the shallow parsing processing would allow only for a mapping onto underspecified semantic modifiers. This produces a loss of information with a a possibly big impact during the generation phase. On the other hand, the corpus texts presents a large array of constructions for which it was really hard to find the appropriate IRep4 equivalent. This has been particularly true for idiomatic constructions and various types of multiword expressions and modifiers. For these correspondences, a transfer component for language pairs would have been helpful.

A classical trade-off between language-independence and depth of linguistic processing, which is known from machine translation, also prevails in MUSI. IRep4 does not contain elements specific to any of the four languages involved, but the analysis results reflected the grouping and nesting of phrases and clauses of the source language. For instance, Italian often uses a post-nominal complex adjectival construction, where German would use a relative clause.

IRep4 also has a few intrinsic deficiencies that affected generation. Most prominently, it does not represent scope and constituent order information. For instance, the scope of negation is important for the placement of the particle, and some order information is relevant to represent the argumentative structure in complex sentences and ensure coherence. Clearly, here another classical trade-off with interlingua becomes apparent: it can only represent what can be analyzed. In MUSI, analyzing scope and thematic structure was not among the tasks set out.

5. System evaluation: preliminary results

As the reference domain for the first phase was the medical domain, validation was carried out by Doctors and Researchers coming from several Hospitals and Institutes of Biology in France and U.S.A. Results cannot be considered as definitive because in this phase only a partial prototype running on a 2000-word lexicon was available. However, the general feeling is rather good and encouraging.

The evaluation method used is inspired by the TIPSTER SUMMAC summarisation evaluation (Mani *et al.* 1998). For a discussion on summarisation evaluation methods, see (Jing *et al.* 1998). Performance of the MUSI summarisation tool was measured in two aspects :

-   **intrinsic quality**. In our project this test is mostly associated to the cross-lingual aspects. As the language of the summary is not the same as the input, it is checked here that the result in the target language is acceptable from the linguistic point of view.

-   **extrinsic quality**. This test deals with the added-value of MUSI summaries in terms of detection of relevance by a human user when integrated in a search engine. In other words, we evaluate how well the MUSI summary helps a person to estimate the relevance of a document with respect to a given subject. To compare with other software, only the monolingual (English-to-English) version of MUSI was used here.

5.1. Intrinsic tests

The intrinsic tests in the English-to-French version have been performed with 15 sentences coming from three articles and 6 evaluators in a blind context. Three different types of summaries have been evaluated in the intrinsic test : human, automatic translation, and MUSI. Automatic translation actually is a mixture of sentence extraction and automatic translation (Reverso commercial software). Notes are on 5 (1 = horrible, 3 = medium, 5 = excellent). The results of the intrinsic evaluation are:

Intelligibility	
Automatic translation	2.00
Human	4.33
MUSI	1.83

Correction	
Automatic translation	2.17
Human	4.33
MUSI	2.33

Fluency	
Automatic translation	3.83
Human	4.50
MUSI	4.83

(Lack of) repetitions	
Automatic translation	3.83
Human	4.50
MUSI	4.83

The average are	
Automatic translation	2,96
Human	4.42
MUSI	3,46

The conclusion is that human texts are better, but MUSI cross-lingual texts are as good as the state of the art in commercial automatic translation. Note however that MUSI is not an automatic translation software, as it simplifies the result: in fact, the goal here is to produce indicative summaries, and not the total fidelity with respect to the source.

5.2. Extrinsic tests

Four articles and three queries have been used. Evaluators were asked to assign relevance to an article after reading its summary. Summaries used were Baseline (first 5 sentences of the source document), Copernic (commercial summarisation software) and MUSI query-oriented (all for 5 sentences) (Once again 1=not relevant at all, 3=medium, 5=central topic).

After this, evaluators were asked to assign relevance after reading the whole articles. The average of these relevance values is taken as the "ideal" reference. To measure the distance between a given summary and the "ideal" reference, the average of the differences between the "ideal" reference and the given summary is calculated. Result are shown in the following table:

	Deviation
Baseline	0.72
Copernic	0.39
MUSI	0.22

It appears that the distance between Baselines summaries and the "ideal" reference is rather high. Among the tested systems, MUSI appears to be the closest to the "ideal" reference. As a conclusion, we can say that we have some evidence of a difference in extrinsic quality between MUSI query-oriented and Copernic. In other words, MUSI query-oriented summaries would be better than Copernic ones.

Of course, the sampling space is too reduced to deduce definitive conclusions by the moment. Only the generalised extrinsic tests (in a second phase of the project) will allow us to be sure of the results. In addition, one must consider that compared to other summarisation software, MUSI offers supplementary functions like cross-lingual summarisation, feedback and adaptability to the domain.

6. References

Busemann, S. (1996), "Best-first surface realization", in Scott, D. (ed.), *Eighth International Natural Language Generation Workshop Proceedings*, Herstmonceux: 101-110.

Busemann, S. and Horacek, H. (1998), "A flexible shallow approach to text generation", in Hovy, E. (ed.) *Nineth International Natural Language Generation Workshop Proceedings*, Niagara-on-the-Lake: 238-247.

Chevreau, K., Coch, J., Cartier, E. (2000), *MUSI Internal Text Representation Specification, Technical* report, MUSI D2.1.

Coch, J. and Chevreau, K. (2001), "Interactive Multilingual Generation" in Gelbukh A. (ed.): *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science Vol. 2004*, Springer Verlag, Berlin: 239-250.

Federici, S., Montemagni, S., Pirrelli, V. (1998) "Chunking Italian: Linguistic and Task-oriented Evaluation", in *Proceedings of the First International Conference on Language resources and Evaluation*, Granada, Spain.

Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. (1998), "Summarization evaluation methods: experiment and analysis", *Working notes of the AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, California.

Lenci A., Montemagni S., Pirrelli V., Soria C. (1999), "FAME: a Functional Annotation Meta-scheme for multi-modal and multi-lingual Parsing Evaluation", in *Proceedings of the ACL-IALL Workshop "Computer-mediated Language Assessment and Evaluation in Natural Language Processing"*, Maryland (USA).

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

Mani, I., Firmin, T., House, D., Chrzanowski, M., Klein, G., Hirschman, L., Sundheim, B., Obrst, L. (1998), "The TIPSTER SUMMAC Text Summarization Evaluation: Final Report", MITRE Technical Report MTR 98W0000138.

Mani I. and Maybury M. (1999), *Advances in Automatic Text Summarization*, Cambridge, MA, MIT Press.

Sparck-Jones, K. (1999), "Automatic Summarizing: Factors and Decisions" in Mani and Maybury (1999).

Tombros A. and Sanderson M. (1998), "Advantages of Query Biased Summaries", in *Information Retrieval*, SIGIR'98, 24-28 August, 1998, Melbourne, Australia.