# Semantic Indexing of Document Bases

Roberto Basili, Maria Teresa Pazienza

Department of Electr. Engineering, University of Rome, Tor Vergata, ITALY
{rbas, pazienza}@info.utovrm.it

**Abstract.** Browsing and navigating into a document base can be significantly improved by an easy access to textual sources. Many efficient indexing and search techniques have been proposed in the literature. Word vectors are commonly used to approximate the notion of document content and to support matching algorithms during the retrieval process. Efficiency criteria push for linear non-recursive representations. The text of a document is never processed for its linguistic information content. The gap between the implicit content of (a set of) texts and the rich structured formats (i.e. networks) able to support intelligent browsing is well known.
In this paper the overall architecture of a language oriented methodology of document processing for a content driven retrieval is described. Lexical acquisition modules are integrated with indexing and browsing ones, in order to support a significant semantic coverage and to guarantee portability throughout different domains. The experience in the development of different IR systems (based on linguistic processing of document content) is used to demonstrate feasibility and strengthness of the methodology.

## 1. Information Search in a Document Base.

Browsing and navigating into a document base can be significantly improved by an easy access to textual sources. Text is still the preferred form for information exchange and storage. Moreover speech processors show us future scenarios where new information is directly provided by voice, in a very unstructured and implicit fashion. In any case the capability of processing language is a fundamental requirement for future information resources interfaces. Their "strengthness" will be also measured according to their flexibility, i.e. integration of different (linguistic vs. non-linguistic) knowledge sources they offer. Our methodology deals mainly with textual information sources. The goal of the paper is to describe a layered environment in which many systems cooperate to realize an intelligent browsing modality. The end user may interact with the overall environment as well as with only one system as his needs change[1].

It is still common opinion that main task of an IR system is to find documents relevant to a user query. Query languages are highly structured formal languages. Further facilities offered to the user are navigation capabilities through the document sets, supported by a global document network. Friendliness and flexibility are highly increased when a hypertextual non-linear organization of the text is implemented by the underlying data structures.
Hypertextual modalities of navigation are widely available for specific application tasks, as in computer-aided formation tools or in educational software. In these applications in fact very narrow objectives inspire the development of the final document hypertextual network. For more general purposes (e.g. research as well as documentation maintenance, circulation and dissemination) criteria for the development and implementation of hypertextual structures are unpredictable.

When querying a document base, it is suitable to rely on an objective notion of relevance. On the contrary, in navigation tools, leaving the user free in selecting the information of interest is an important aspect. The modality of individual browsing is highly subjective. Objectivity and subjectivity of the search mode are not independent from the content and structure of the underlying textual information. In order to define an objective notion of relevance a clear semantics of the underlying communicative act must be provided. These two aspects are clearly related to the linguistic information that is (implicit) in text.

The implicit content of a document is made of two different components. The first is related to the proper linguistic information, i.e. the set of (coordinated) sentences that form the communicative act of the writer. The second one, totally independent from the first, is related to the structure of the set. Sections, chapters, footnotes contribute to improve the informative

[1] In the paper we shall frequently refer to our previous publications in which details on the architectures and methodologies of the different systems are provided. Hereafter we shall point out a new browsing modality into a document base that exploits linguistic information related to texts as well as structural information fitting the form of documents.

content of the resulting message, according to highly variegated and domain dependent (implicit) rules. Generally this structural information (e.g. a title) is able to deeply characterize the content of a document . Sometimes part of this structural information (e.g. the author field) becomes an essential part of the content itself, able to justify a specific browsing. Any knowledge-based interface should rely in some explicit or implicit way on a clear distinctions of these two separate levels. An explicit distinction is realized in systems where a portion of the knowledge is devoted to fully declare "what is" a document, its qualified structure and what parts can efficiently support a specific search. Other systems may simply rely on a separation implied by independent control functions acting on different document portions. In these last cases the two levels are only implicitly distinct.

From the indexing point of view the distinction between these two aspects reflects the more general duality objectivity/subjectivity.
The textual level in fact refers to the implicit (and generalized) knowledge shared by the whole user community of a language. The text component of a document may be then considered as an objective *property* of a document, independently from the interest in accessing it by an end user. Whatever is the application task, the focus of a scientific abstract about infra-red sensors, for example, is always related to electromechanical devices used for Remote Sensing even if the current reader is a mathematician interested to differential equations.

The structural information is more specific (and often application dependent): it refers to the way the documents parts are accessed, so that it is not an objective property of a document. In a classification system, for example, the author and the title may be of primary importance. In an hypertextual system, for educational purposes, the authorship may be irrelevant, and thus be removed from the document storage format.

In order to fill the gap between the information implicit in a text and its structured form suitable to support browsing and navigation, a knowledge-based interface should be designed to manage both aspects: structure and content.

This paper deals with an application environment in which three different systems cooperate: ARIOSTO (see Appendix 1) for linguistic processing and extraction of semantic information from texts; HIRMA for document semantic indexing within a vector space model and natural language query support; HERMES for hypertext specification and browsing.

In our integrated environment the end user is allowed to *query* (in natural language) a document base, to *specify linking criteria* and *browsing* within a resulting document network.

## 2. Document Semantic indexing

Knowledge-based interfaces to information resources, whatever is the supported search modality, should rely on expressive and efficient representation forms. Linear structures, as word vectors, are generally preferred as they easily generalize in metric spaces. Further generalizations associate frequency based weights to vector components. However the *relevance of a word* (or phrase) *in a text is independent either by its frequency or by the text size*. For example in a domain of Remote Sensing scientific abstracts we acquired evidences that the information content of the word *satellite* should not be shared among different word senses. Sentences like "*the satellite flies over ....*" and "*Earth satellite (i.e. moon) ...*" cannot be taken as separate, without any representation level different from string cooccurrencies. In sentences like " *Earth Observation systems ....*" the notion of *satellitary platform* is expressed in a different fashion. Capturing such equivalence is another important target. In fact, word senses represent more of the semantics of the text as they provide a basis for exploring lexical semantic relationships (such as synonymy and antonyms) (Krovetz and Croft,1992).

*Word senses* should be an essential component of any document representation language. Any such semantic representation implies that similarity among senses results in similarity among documents. When a document base is described by means of word senses rather than words, retrieval performance may be consequently improved.
Senses should not be too general, to appropriately reveal the knowledge content of each document. In fact, senses distribute in a space that is not only dependent on the representative of the language itself but of the described world, (i.e. the specific domain ontology). Furthermore, a user relates senses to words in the query to express his thinking activity. More important senses are strongly related to the topics the user community is interested in. User interest focuses on classes of arguments relevant to the knowledge domain. Symptoms, Medicines or Pathologies are qualitative examples of such classes within a (possible) medical domain. Word senses contribute thus to identify concepts in the document base that are important for the users.

Examples of such meaningful concepts are the word sense of *satellite* or *IOL* in sentences like

*"Earth observation satellites for future applications employ high resolution sensors (...)"*
*"Especially for the transmission of such high data rates on IOLs and ISLs optical systems are (...)"*

These different words behaviors (i.e. content) are defined as word *contextual roles* in our system, (Basili et al., forthcoming).The set of contextual roles relevant for the document base is derived by the analysis of the related sublanguage. Contextual roles are detected in texts and expressed by suitable semantic primitives called *representative types*. Roles and lexical expectations are also central to other text processing systems, like SCISOR (Rau, Jacobs 88). Technically, contextual roles are domain dependent conceptual graphs *schemata* (Sowa 84,88) that are possibly filled by relevant sentences and passages of the documents. Indexing is realized by triggering schemata in an expectation driven fashion. An example of schema defined in the Remote Sensing domain is the following:

```
(1)  Remote_Sensing_Machinery(x)  iff
       [INSTRUMENTALITY:*x]
            <-(INSTRUMENT)-[ACT:*z]
                  -(OBJECT)->[LOCATION:*y]
```

This definition allows the system to assign the role `Remote_Sensing_Machinery` to the word *system* in *"Earth observation systems ..."* as well as to *Sar* in *" ... ERS-1 Sar enlights land surface with a resolution of 30 m"*.

The semantic interpretation of a given document produces the list of the corresponding relevant words, tagged by the detected contextual roles. The couple `<word,contextual role>` is a potential anchor in an hypertextual search modality as well as a semantic component of a document in a vector space model. ( Details of document semantic indexing in the vector space may be found in (Marega et al.,1994)).Both these representations have been exploited in HERMES (Basili et al.,1994), and HIRMA (Basili et al., forthcoming) systems (access to these references for details). As an example in the figure 1 has been shown the text of a document (an abstract of a scientific paper in Remote Sensing domain) and its semantic vector, that is the result of the linguistic processing of the text by ARIOSTO. The last one is a list of couples <Contextual role, Word> and their occurrences in the document. A word sense is thus each component of the vector; infact a word may be used in the text with different senses (as, for example, *rate* in figure).

```
      Optical data communication for Earth observation satellite systems
      Future Earth observation satellite systems for worldwide high
    resolution observation purposes require satellites in low Earth
    orbits, supplemented by geostationary relay satellites to ensure
    intermediate data transmission from LEO to ground. Earth observation
    satellites for future applications employ high resolution sensors and
    deliver data rates of at least 500 Mbit/s. Especially for the
    transmission of such high data rates on IOLs and ISLs optical systems
    are well suited. (...)

Semantic Vector:

Observed_area                  earth          1
Experiment_activity            observation    1
R_S_machinery                  satellite      3
R_S_machinery                  system         1
Experiment_activity            transmission   1
Focus                          rate           1
Kind_of_processed_information  rate           2
Utility_machinery              iols           1
Utility_machinery              isls           1
```

**- Fig. 1: A source document and the related semantic vector -**

## 3. Browsing

Hypertextual navigation is the browsing method used by HIRMA and HERMES systems. We may think of an hypertext as a nonsequential text, that is a text consisting of chunks of information (*documents*) and cross-references among them (*links*). Links do not connect whole documents, but single parts of them. We will call *anchors*

those document parts that are starting point or end for a link. An anchor will represent a significant descriptor of a document; it may refer both to its structure and content. The document structure by itself is able to characterize different document types (such as abstracts vs. encyclopedia's entries or full documents...) and may suggest links between related substructures. A document type may be defined as an n-uple of attributes that constitute its components. The internal reports of a factory, for example, are characterized by a given structure: header, division, keywords, authors, date,... Each one of these attributes may be crucial in establishing links to other documents. As an aexample, documents of the same division (or of strictly dependent ones) may be highly related (as a default) and may require a connection via a link.
Every attribute ha a type; attribute type may be atomic (such as STRINGs, NUMBERs, LABELs), structured (i.e. a n-uple of ordered attributes) or finite sets(i.e. any aggregate of values of the same type).
TEXT is a primitive structured attribute describing a portion of texts. Among the attributes defining TEXT, ANCHORS is a set type of the contextual anchors related to the text. The anchor is characterized by the word, its location in the text and and the contextual role assigned to it. For example in the sentence "Earth Observation satellite system..." *satellite* is an anchor as it can be assigned with the contextual role of 'R.S. Machinery'. The number of anchors is independent from document length and expresses the richness of different information it contains. The set of available documents and links will be hereafter referred as forming the *document base*. The hypertextual browsing module includes the tools for creating, updating and navigating into the document base. HIRMA (Basili et al., forthcoming) allows the integration of querying and browsing. Navigation in fact, can be triggered by documents relevant to a NL user query.
HERMES- is the browsing system. It consists of different modules:

- the **Schema Definition module** accesses an high-level specification of the hypertext (i.e. the schema definition file ) , and produces the updating module. The logic specification of the hypertext is compiled by the hypertext manager[2] ;
- the **Pre-Processing module** extracts the semantic representation and the structure of a document. It takes raw texts files as input and returns structured (i.e. interpreted) documents ;
- the **Updating module** integrates new documents into the document base. The upgrade may be performed at different phases of the hypertext life-time (creation as well as run time);
- the **Browsing interface** allows the user to navigate through the document base. This interface can be also activated by the IR system, at the end of the retrieval process.

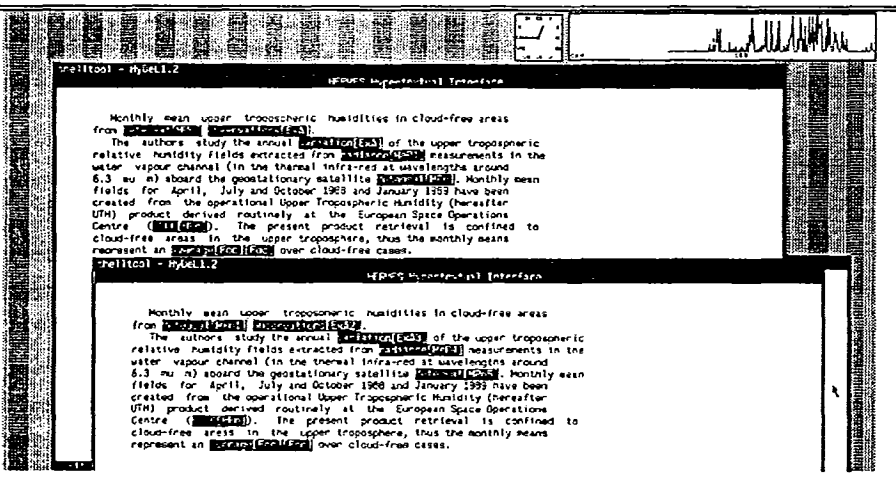
In HERMES the activities involving human intervention are only related to *the definition of contextual roles* that rules the document interpretation and *the compilation of the schema definition*. Hypertext schema definition does not require specialized skill but just some knowledge about the documents. After the document preprocessing, the detection of useful links and the generation of the proper navigation libraries can be fully automatized. The hypertext schema is the formal specification of the links of interest for the user. It is provided by a dedicated language, that we will refer as the *hypertext definition language* (HyDeL, (Basili et al., forthcoming)).

### 3.1 Authoring

The *authoring* activity is the hypertext technology bottleneck. It is generally performed by the hypertext manager, who knows the hypertext target users, reads *every* document and decides *subjectively* what links what. This is an expensive, not error free job.
In HERMES, the hypertext manager assumes the charge of specifying *link validation rules* for the system. The required knowledge refers to the source document formats and a dedicated rule specification language used by the end user to declare the requirement to be satisfied by the documents to be inserted in the hypertext.
This reduces the human activity of hypertext maintenance to the development of such rules that form the Ht schema definition. In such a schema all the information on the target hypertext, i.e. rules and criteria for linking documents as well as text structures, are provided by using HyDeL. Any other operation is committed to the system that compiles the rule specifications and automatically generates the updating libraries. The automatic generation of links is rule-based and this favours a clear semantics of the underlying hypertextual structures. This is an essential improvement in user browsing.

[2] The hypertext manager is the responsible for document definition and link organization, the new insertions, the updating as well as browsing modalities.

*Anchor extraction* refers to the identified meaning of the document and to its structure: then anchors are a property of a document and are independent from *links* that are activated after an hypertext definition. In fact they pre-exist to link generation and continue to be valid independently from the insertion of the document in a hypertext. In such framework it is possible to define different hypertexts on the same set of documents. As an example in fig. 3, it is shown the same document as it appears in two different hypertexts. Within this environment different views of the document base can be supported. In fact, whenever the system manager is the only responsible for the hypertext definition, only selected components of the document base may be accessible by the end user.



Legenda:
*Bold-faced typed words point out anchors (that is ouple word-contextual role). Some of them are activated and become source or destination of a link, according to the hypertext definition. The right side number expresses the active status of the anchor.*
*It is evident that two different hypertexts are here depicted, as different anchors are active.*

- Fig. 3: Multiple sessions in Hermes -

Due to fact that the anchor detection is indepenedent from the hypertext generation phase, the updating process my be carried on incrementally. The document description is in fact always valid (i.e. objective) and no document reprocessing is necessary: the new texts will be integrated with the old ones just according to their representation (i.e. set of anchors).

HyDeL is the language used to specify the hypertext schema definition. Schema definition includes rules that express classes of links to be instanciated within documents. The main sections of the schema definition includes:

- a *document type declaration* (that's any kind of attribute, and their type; even structured attribute have to be defined here);
- a *collection definitions* (collections are set of documents, stored together for any reason; a collection can be visited as a document, can be source or destination of links);
- a *link definition* (link rules express criteria for linking two documents).

Let us show a typical example of rules, as they have been defined in the Remote Sensing domain where both structured (DIF[3]) and unstructured (abstracts) documents have been collected. The rules may be qualitatively expressed by

*(R1) Activate a link between two DIF documents if their field DISCIPLINE is the same*

*(R2) Activate a link between an abstracts and a DIF document iff a contextual anchor of the abstract whose label is 'Measured Parameter' and the field GEOPHYSICAL PARAMETER of the DIF document are the same.*

Rules *(R1)* and *(R2)* are specified in HyDeL as follows:

```
(R1)  define L1 between
        X(STRING) in D1/DIF_DOC.dislist,
        Y(STRING) in D2/DIF_DOC.dislist
      where < X/STRING = Y/STRING >.
```

3 DIF documents are structured documents according to the NASA Directory Interchange Format specification (Nasa,88). They are part of the Remote Sensing Corpus of documents managed by HIRMA, and HERMES.

```
(R2)  define L2 between
        X(ANCHOR) in D1/ABSTRACT.anchors,
        Y(STRING)  in D2/DIF_DOC.parlist
      where < X/ANCHOR.word = Y/STRING >.
```

In (R1) ad (R2), L1 and L2 are the different links to be activated and D1 and D2 are instanciated by the documents to be compared.

## 4. Discussion

The environment that have been mentioned (and very summarized) in the previous sections are:ARIOSTO (Basili et al.,1992a,b,1993a,b1995), a corpus driven system for lexical acquisition that semiautomatically derives a semantic lexicon in an application sublanguage by means of logical as well as statistical techniques; HIRMA(Basili et al.,forthcoming), an Information Retrieval system making use of natural language, as a query language, and semantic indexing algorithms; and HERMES (Basili et al.,forthcoming), an hypertextual navigation tool based on natural language processing techniques to generate and browse a document base. The overall environment here depicted has been experimented on a collection of Remote Sensing documents. The lexical knowledge base for HIRMA has been acquired from the corpus (about 700.000 words). The source language is English and the linguistic processing make use of a(n automatically acquired) semantic lexicon of some hundreds of word senses. Specific grammars and morphologic dictionaries for English have been adopted. A running version for the italian document bases exists: it is supported by dedicated language dependent tools.

Tests have been carried on three collections: the mentioned Remote Sensing collection, the collection of Italian document on V.A.T. laws (about 450.000 words) and an equivalent collection of summaries of newspaper articles on ecology and the environment (about 350.000 words). The generated document bases for HERMES range from some hundreds to few thousands documents. The hypertext connectivity has been measured under reasonable thresholds: no exponential growth in the number of (automatically generated) links is observed (Basili et al., 1994).

Some aspects of the overall methodology are worth to be stressed. The exploitation of the incoming textual sources provides an intrinsic data-driveness that should be highly encouraged for truly user-oriented systems. The used acquisition tools imply also a significant domain specificity of the terminology and of the lexical knowledge (i.e. the semantic lexicon and the grammars). This favors an high portability: the experience of porting the systems to different languages and domains (e.g. V.A.T. laws vs. ecology) demonstrated the economy of the overall strategy. A small team of two-three people is able to adapt the lexical knowledge base to a new domain in some weeks. The different implemented IR shells demonstrated a substantial feasibility in medium sized domains (500-2000 documents).

Lexical analysis demonstrated that a significant coverage has been obtained: the shallow semantic representation (i.e. canonical graphs and schemata for index terms) is very specific and the results are meaningful. A study in the ecology domain showed that only a small percentage of documents (5-6 %) results in unsatisfactory indexes. Further customization is also possible within the acquisition framework of the ARIOSTO system (Basili et al.,1993). The use of natural language as a query language allows possible developments for discourse analysis and more complex forms of interaction.

An essential feature of the methodology is the particular notion of authorship, that NLP techniques provide in HERMES. In HERMES, the hypertext generation phase is separate from the document analysis and interpretation. Several hypertexts can rely on the same document base.

The development of a special purpose hypertext definition language (i.e. HydDel) provide different users imposing *different views of the documents,* with different (customed) underlying networks.

Anchors (objective properties) are document features, while links (subjective view) are (strictly) user defined. Subjectivity can be preserved for browsing purposes, while the logical representation of the documents is shared over the same document set. It is in fact the common knowledge about the content (i.e. conceptual graphs) that can possibly be exploited by further intelligent activity (e.g. document maintenance, hybrid graphical and textual reasoning, ...). The methodology provides thus something more than browsing or querying support.

Further work should be done on methodologies of *supervision*. The system actually is not supervised when proposing and generating document links. Rules are compiled by hand for significant sentence frames and then applied to all the documents. The study of the best levels where to apply human validation and control is still on going. Furthermore, user models, in

terms of user dependent document relevance or weighted links, should be introduced.

## 5. References


(Basili et al. 1992a) R. Basili, M.T. Pazienza, P. Velardi, A shallow syntactic analyzer to extract word associations from corpora, Literary and Linguistic Computing, 1992, vol. 7, n. 2, 114-124.

(Basili et al. 1992b) R. Basili, M.T. Pazienza, P. Velardi: "Computational Lexicon: the Neat Examples and the Odd Exemplars", Proc. of 3rd Conference on Applied NLP, 1992.

(Basili et al. 1993a) R. Basili, M.T. Pazienza, P. Velardi, Acquisition of selectional patterns, Journal of Machine Translation, 8:175-201,1993.

(Basili et al. 1993b) R. Basili, M.T. Pazienza, P. Velardi: *"What can be Learned from Raw Text? An Integrated Tool for the Acquisition of Case Rules, Taxonomic Relations and Disambiguation Criteria"*, Journal of Machine Traslation, 8:147-173,1993.

(Basili et al., 1994), R. Basili, F. Grisoli, M.T. Pazienza, *"Might a Semantic Lexicon supports Hypertextual Authoring?"*, Proc. of the Applied Natural Language Processing Conference, Stuttgart, October 1994.

(Basili et al., 1995) Basili, R., Pazienza, M.T., Velardi P., *A context driven conceptual clustering method for verb classification*, in B. Boguraev, J. Pustejovsky Eds., MIT Press, in press.

(Basili et al., forthcoming), R. Basili, F. Grisoli, M.T. Pazienza, *H.I.R.M.A. Hypertextual Information Retrieval System Managed by ARIOSTO*, to appear on Data and Knowledge Engineering Journal.

(Dahl 89) V. Dahl: "Discontinuous Grammars", Computational Intelligence, Vol. 5, pp.161-179, 1989.

(Krovetz and Croft,1992) R. Krovetz, and B. Croft, Lexical Ambiguity and Information Retrieval, ACM Transaction on Information Systems, vol;. 10, n. 2, April 1992, pp. 115-141.

(Marega, Pazienza 94) R. Marega, M.T. Pazienza: "Co.D.H.I.R.: an Informaion Retrival System Based on Semantic Documnt Representation", Journal of Information Science, vol.20,n.6,1994.

(NASA 88) NSSDC: "DIF: Directory Interchange Format Manual", Version 1.0, NASA, Goddard Space Flight Center, July 13, 1988.

(Rau, Jacobs 88) L. F. Rau, P. S. Jacobs, Integrating Top-Down and Bottom-Up strategies in a Text Processing system, Proc. of the Second Applied Natural Language Processing, 1988, MorrisTown, NJ, 1988.

(Sowa 84) J. F. Sowa: "Conceptual Structures: Information Processing in Mind and Machine", Addison-Wesley, 1984.

(Sowa 88) J. F. Sowa: "Using a Lexicon of Canonical Graphs in a Semantic Interpreter", in Relational Models of the Lexicon, M.Evens, Ed., Cambridge University Press, 1988.


## Appendix 1: ARIOSTO: system overview

ARIOSTO[4] is an integrated system for the acquisition of large-scale lexical information and computational lexicon, based on a mixed approach: both NLP metodologies and statistical techniques are used.

The core lexical knowledge acquired by ARIOSTO is based on word association data augmented with syntactic and semantic markers. For this type of analysis a complete syntactic processing is unnecessary, because what matters is the detection of all binary or at most ternary relations among content words rather than among more complex constituents. In ARIOSTO statistical tools are applied to syntactic information extracted from corpora. After a morphologic analysis the syntactic analysis is performed by a simplified parser, based on a discontinous grammar formalism [Dahl 89], which is able to detect binary and ternary syntactic relations between words, that we call *elementary syntactic links* (esl). The use of a discontinuos grammar, instead of statistical techniques, (e.g. windowing), allows better linguistic performances (accuracy and precision) without loosing in computational efficiency.

The most part of the knowledge is then acquired by ARIOSTO directly from corpora, with statistical tools; human work is confined to the realization of preprocessing tools. ARIOSTO requires work of experts only for the *definition* of:

- a morphologic lexicon of terms,
- grammars (a generative grammar for the morphologic analysis and a discontinuous grammar for the syntactic analysis),
- a taxonomic knowledge base (i.e. an high level semantic tagging of words and semantic relations).

The very general approach and the processing modularity achieve an effective portability among different logic programming styles, linguistic domains (currently legal, commercial, environmental, remote sensing, are processed) and languages (both Italian and English versions have been implemented).

A very meaningful result of ARIOSTO is the production of a specialized domain-dependent lexicon for each corpus.

The size of our processed corpora are approximately 500,000 words each, thus ensuring homogenous source data, awareness of slang words behaviours, dominance of similar approaches, all that cooperate in a significant meaningful description of words The linguistic and statistical processing realized by ARIOSTO are carried out only once for each corpus. Results are then stored and used later for different applications (IR and Ht are only two of them).

[4] ARIOSTO exploits part of the experience acquired within previous systems DANTE and PETRARCA based purely on NLP techniques and developed by our research group.