The MEANING Italian Corpus

Luisa Bentivogli, Christian Girardi, Emanuele Pianta
ITC-irst
Via Sommarive 18 - 38050 Povo (Trento) – Italy
E-mail: {bentivo, cgirardi, pianta}@itc.it

Abstract
The MEANING Italian Corpus (MIC) is a large size corpus of written contemporary Italian, which is being created at ITC-irst, in the framework of the EU-funded MEANING project. Its novelty consists in the fact that domain-representativeness has been chosen as the fundamental criterion for the selection of the texts to be included in the corpus. A core set of 42 basic domains, broadly representative of all the branches of knowledge, has been chosen to be represented in the corpus. The MEANING Italian corpus will be encoded using XML and taking into account, whenever possible according to the requirements of our NLP applications, the XML version of the Corpus Encoding Standard (XCES) and the new standard ISO/TC 37/SC 4 for language resources. A multi-level annotation is planned in order to encode seven different kinds of information: orthographic features, the structure of the text, morphosyntactic information, multiwords, syntactic information, named entities, and word senses.

1. Introduction

A domain-based corpus can be a useful resource in different research areas. It is well known that domain-specific sublanguages exhibit specific features at various linguistic levels (Grishman and Kittredge, 1986). Linguistic analyses carried out on a multi-domain corpus can uncover differences in the lexicon and morphology, in names and named entity structures, and in lexical semantics, syntactic and discourse structure. The NLP community can find in a domain-based corpus a fundamental resource for several tasks such as, for example, parsing (Sekine, 1997), domain-dependent lexical acquisition, Word Sense Disambiguation (WSD), etc. In WSD, domain lexical information has proved to be very useful in the development of high precision algorithms (Magnini et al., 2003).

Domain (also called *topic*, or *subject matter*) is one of the criteria for text selection and/or classification in many existing corpora. For instance, in the written component of the British National Corpus two main text selection criteria were used: "medium" and "domain". More specifically, the BNC uses 9 knowledge domains (arts, social sciences, world affairs, etc.). Similar selection criteria (medium and domain) have been adopted also in the design of the American National Corpus (Ide and Macleod, 2001). The Brown and LOB corpora classify texts in 15 different text categories but such categories are a mix between genre labels (bibliography, popular lore) and domain labels (religion, "skill, trade and hobbies"). The NERC report (Calzolari et al., 1995), offers a summary of the classification systems used by major corpus projects in Europe, showing that domains are generally used in the classification of the texts. The same holds for the most important corpora created for the Italian language. In the SI-TAL Italian Treebank (Montemagni et al., 2000) texts have been "selected to cover a good variety of topics". The reference corpus CORIS/CODI (Rossini Favretti et al., 2001) is structured in subsections, some of which can be compared to domains.

However, in all the mentioned corpora, a complete representation of domains is not pursued in a systematic way. On the contrary, domain is the fundamental selection criterion of texts to be included in the MEANING Italian corpus (MIC). The MIC is being developed with the aim of supporting domain-based WSD in the framework of the MEANING project (Rigau et al., 2002). MEANING is an EU funded project which aims at enriching existing wordnets (for English, Spanish, Catalan, Basque, and Italian) by acquiring new lexical information from corpora. MEANING tries to exploit the inter-dependency between Word Sense Disambiguation (WSD) and knowledge acquisition by applying the following steps: *1*. Train accurate WSD systems and apply them to very large corpora; *2*. Use the partly disambiguated data in conjunction with shallow parsing techniques and domain information to extract new linguistic knowledge to be incorporated into wordnets; *3*. Re-train WSD systems and re-tag the corpus, exploiting the information acquired in the second step. The result of this cycle is twofold: the enrichment of the lexical resources with information acquired from the corpus and a multi-level linguistic annotation of the corpus itself.

The rest if this paper is structured as follows. In Section 2 the structure of the MIC is described in detail. Section 3 deals with the encoding of the corpus while in Section 4 the multi-level linguistic annotation of the corpus is illustrated with annotation scheme examples. Section 5 summarizes what has been done up to now and what are the tasks still to be undertaken.

2. Corpus design

The MIC is being created with the aim of representing the domains used in WORDNET DOMAINS (Magnini and Cavaglià, 2000), an extension of WordNet 1.6 where each synset has been annotated with at least one domain label, selected from a set of 164 labels hierarchically organized. The WORDNET DOMAINS hierarchy was created starting from the subject field codes used by current dictionaries, and the Dewey Decimal Classification system (DDC), a general knowledge organization tool that is continuously revised to keep pace with knowledge development. The DDC is the most widely used library classification system in the world and provides a very large and complete set of hierarchically structured domain labels (see DDC 1996). A core set of *42 basic domains* (the second level of the WORDNET DOMAINS hierarchy) has been chosen to be represented in the MIC. A study carried out by (Magnini and Gliozzo, 2002) shows that these 42 domains have a domain-coverage equivalent to the domain-coverage of the DDC system.

In the MIC, texts are assigned to a topic category on the basis of an existing, *text-external*, list of domain labels. The value of this kind of classification is one of the central controversial areas of text typology, as pointed in the EAGLES preliminary recommendations on text typology (Sinclair and Ball, 1996). This report argues that it is not possible to classify the texts produced in the world on the basis of a limited list of topics, chosen on a text-external, a priori basis; there are too many possible methods for identifying the topic of a text. Also, the boundaries between topics are blurred, and texts usually cover a variety of topics. On the contrary, the topic(s) of a text should be identified on the basis of *text-internal* evidence such as vocabulary clustering.

However, while claiming that internal evidence should be the primary criterion for the identification of a text topic, the EAGLES report admits the possibility of a defensible use of topic categories based on few external criteria. These are the sectionalisation of newspapers, some topic-related classifications institutionalized in a society (in particular lists of recognised professions and educational courses), and, when existing, the self-classification of the text.

We recognize the problems mentioned above and we agree with the position that there is no objective, scientific means of assigning topics. However, a commonly accepted topic classification scheme based on internal criteria has not been developed yet. Moreover, some practical consideration must be made. In the current corpus practice text-external criteria are widely used to assign topics to texts. As it is shown in the introduction, the topic categories given in the NERC report have a common ground in many or most of the corpora studied. The MIC is in line with the trend in corpus practice as most of the commonly used topics reported in that document correspond to our basic domains. Moreover, as we will see below, in the construction of the corpus we exploit all the acceptable external criteria mentioned in the EAGLES report.

Coming back to the model designed for the creation of the MIC, we were faced with two requirements. First, we needed completeness, i.e. we wanted all of the 42 domains to be represented. Second, we wanted the corpus to reflect the fact that different domains do not have the same relevance in the language. To meet the completeness requirement we are creating a *micro-balanced corpus* composed of 42 subcorpora, each representing a basic domain. On the other hand we are creating *a macro-balanced corpus*, i.e. a homogeneous corpus of the contemporary Italian language created without taking into account the domain criterion but in which we know that most domains are represented. This corpus will allow us to verify in an independent way the relevance of the different domains in the generic language. Figure 1 shows the overall structure of the MIC.

As regards the other corpus building criteria, the MIC represents only the written (electronic) mode, relying on written texts already available in electronic form. The media used are essentially three: newspapers, press agency news, and web documents. The genre is mainly that of informative, "factual" prose.

An important characteristic of the corpus is that a part of it is bilingual. It includes 5 million words of aligned parallel English/Italian news and the first version of MultiSemCor (Bentivogli and Pianta, 2000), which is a bilingual aligned parallel corpus semantically tagged with a shared inventory of senses. Up to now MultiSemCor consists of 30 English texts of the SemCor corpus (a subsection of the Brown corpus semantically tagged with WordNet senses) along with their Italian translations, for a total of about 120,000 words.

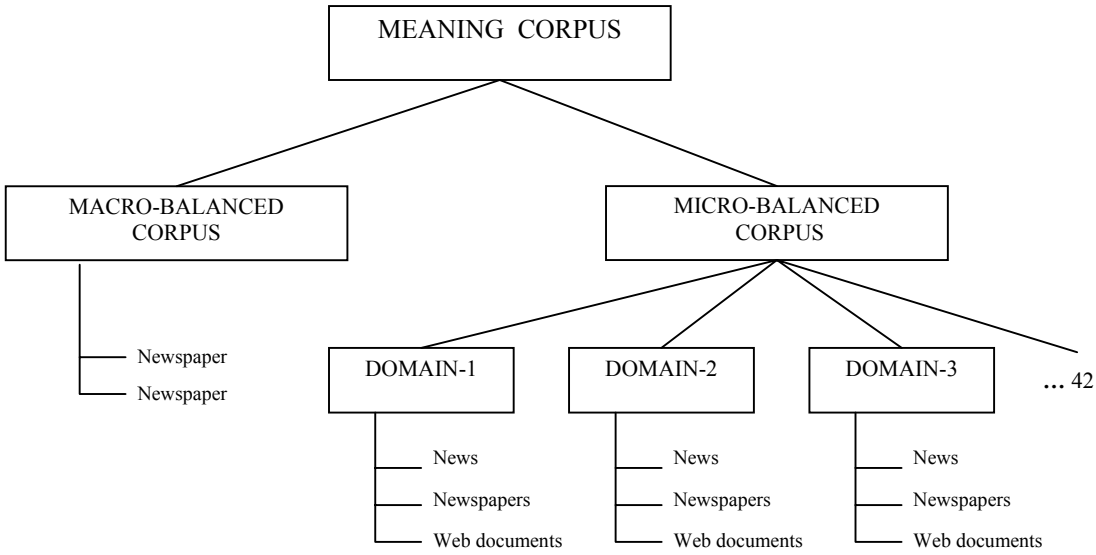

Figure 1 *Corpus Composition*

2.1 The micro-balanced component

The micro-balanced section of the MIC will be composed of 42 subcorpora representing the 42 basic domains selected from WORDNET DOMAINS (reported in Table 1). To create the subcorpora, we take into account the whole hierarchy of WORDNET DOMAINS. This means that for each subcorpus we look for texts belonging not only to the corresponding basic-level domain, but also to the more specific domains related to it in the hierarchy. It is important to underline that the micro-balanced corpus is not composed of specialized texts as we do not aim at creating specialized corpora but a general language corpus in which all the domains are covered. Given the fact that the 42 basic domains seem to have a different absolute relevance, we distinguish *major* domains (e.g. Economy and Sport) and *minor* domains (e.g. Linguistics and Astronomy). Each major domain subcorpus will include 2 million words while the minor subcorpora will be composed of 1 million words each.

The texts to be included in the micro-balanced corpus come from three main sources: press agency news, newspaper weekly special supplements, and web documents. Each domain subcorpus should be balanced with respect to the three media; however for some subcorpora most of the texts will be web documents as it is unlikely that we will be able to find enough news or supplements belonging to those domains (see for instance Mathematics, Pedagogy, Archeology).

2.1.1 Press agency news and special supplements

The press agency news were collected through the Excite (http://www.excite.it) and Virgilio (http://www.virgilio.it) portals. They come from the following press agencies: Reuters, ANSA, ASCA, DataSport, and ADNKRONOS (parallel Italian/English). Supplements come from a wide circulation newspaper called "La Stampa", which contains weekly special supplements dealing with science ("Tuttoscienze"), books ("Tuttolibri"), finance ("Tuttosoldi"), television ("TV"), cars and motorbikes ("Speciale motori"), agriculture ("Speciale agricoltura"), Italian elections ("Speciale elezioni") and local events in the town of Turin ("Speciale città" and "Torinosette").

To speed up the creation of the micro-balanced corpus, we explored the possibility of developing a methodology for the (semi-)automatic classification of news and special supplements to be assigned to the various subcorpora. This is made easier by the fact that both press agency news and special

Administration	Artisanship	Computer science	Law	Philosophy	Sexuality
Agriculture	Astrology	Earth	Linguistics	Physics	Sociology
Alimentation	Astronomy	Engineering	Literature	Play	Sports
Anthropology	Biology	Economy	Mathematics	Politics	Telecommunications
Archaeology	Body care	Fashion	Medicine	Psychology	Tourism
Architecture	Chemistry	History	Military	Publishing	Transport
Art	Commerce	Industry	Pedagogy	Religion	Veterinary

Table 1 *Basic domains in WordNet Domains - Version 1.1*

supplements are already classified by the publishers with two different kinds of information: *broad topic* and *keywords*. News are divided into 9 broad topics, namely economy, politics, cars and motorbikes, artistic performances, sports, science and technology, generic news, foreign news, local news. Also the 9 special supplements can be considered self-classified in 9 broad topics given in their title (science, finance, etc.). Moreover, one or more keywords, often corresponding to domain labels, are always associated to each piece of news and supplement article.

We studied how to develop a procedure able to exploit this information to (semi-)automatically assign news and supplements to the appropriate domain. To develop and evaluate the procedure, a development and a test set have been created for both news and supplements. The *development set* is composed of all the 20,399 news collected from the Excite portal in five months, from April to August 2002. As the news constitute an always growing open set, it is important to verify the productivity of the procedure when applied to news belonging to a different period of time. For this reason, the time-span covered by the test set was kept different from that of the development set. The *test set* was created by selecting 500 out of 15,014 news, collected in five months from September 2002 to January 2003. The news in the test set were chosen randomly, keeping temporal distribution and the proportions of the broad topics with which they were classified by Excite.

As regards supplements, we had at our disposal newspapers special supplements covering a time-span of 10 years (from 1992 to 2001), for a total of 66,927 articles. As these supplements represent a closed set, both the development and test set can be selected from the same period of time. We selected 30,000 articles for the development set and 500 for the test set, randomly chosen but homogeneously distributed over the 10 years and keeping the proportions of each broad topic.

The news composing the two test sets were read and manually assigned to the appropriate domain. Note that the texts have been assigned to the most specific domain available among the WORDNET DOMAINS.

As a first step, we tried a simple algorithm that can be considered as a baseline for our experiment. We associated to each domain a set of Italian words currently used to refer to that domain. This set of words (*domain word set*) has been manually created and contains the lemma of the domain, possible morphological variants, and possible synonyms. As an example, the following is the domain word set associated to the *pharmacy* domain:

Domain:	PHARMACY
Domain word set:	farmacia (pharmacy), farmaceutica (pharmaceutics), farmacologia (pharmacology)

Then, for each domain a procedure looks for a matching between words in the domain word set and keywords associated to the texts. This procedure assigns a text (piece of news or article) to a domain if at least one of the words contained in the domain word set corresponds to one of the keywords associated to the text. The procedure exploits information about keywords as their granularity is similar to that of domains. The 9 broad topics are too generic to be useful for the baseline algorithm.This procedure relies entirely on a priori information, as it does not require any kind of analysis of the development set

The results of the application of the baseline procedure to the test sets of the news and the special supplements are shown in Table 2. In the evaluation, the domain assigned by the procedure and the manually assigned domain are considered to match if they are equal or if they have a common ancestor that is a basic domain.

Since these results show -especially for the news- a very law recall, a second procedure has been developed, based on a number of rules manually written on the basis of the study of the development set. These rules exploit wider information than the baseline algorithm:

- keywords that are not in the domain word set, but are somehow related to the domain
- the broad domains
- the words in the title

Table 3 shows three sample rules, which apply to the *pharmacy* and the *computer science* domains. The first rule only considers information about the text keyword(s). The second rule looks at both keywords and words in the title. The last one considers both keywords and broad topic.

	Precision	Recall	Coverage
News	0.72	0.15	0.20
Special supplements	0.54	0.56	0.70

Table 2 *Performances of the baseline procedure on news and supplements*

PHARMACY	*if*	KEYWORD = farmacia (pharmacy) *or* farmaceutica (pharmaceutics) *or* farmacologia (pharmacology) *or* farmaco (medicine) *or* vaccine (vaccine)
PHARMACY	*if* *and*	KEYWORD = epatite (hepatitis) *or* morbillo (measles) *or* meningite (meningitis) *or* antipolio (polio) *or* virus (virus) TITLE = vaccine (vaccine)
COMPUTER SCIENCE	*if* *and*	KEYWORD = internet (internet) BROAD TOPIC = tecnologia e scienza (science and technology)

Table 3 *Examples of rules*

This second procedure has been developed only for the press agency news and gives the results shown in Table 4 below. The precision decreases but both recall and coverage improve significantly. Unfortunately, despite these improvements, the precision of both algorithms is still unsufficient to avoid manual intervention. Thus, we plan to use the results of the application of the second algorithm for supporting humans in creating the 42 subcorpora. Manual work will be speeded up as corpus builders will have to check if the assignment of the text to a domain is correct or not, a task which is much simpler than assigning a text to one of the 42 subcorpora.

In order to test the applicability of the hand-written rules -developed for the press agency news- also to other texts, we applied them also to the articles of the special supplements. The application of the second procedure to the supplements does not change significantly the results obtained with the baseline algorithm: precision goes from 0.54 to 0.53, and recall from 0.56 to 0.57. These results demonstrate that the rules created on the basis of the press agency news are specific to the news themselves and cannot be reused for different kinds of texts.

2.1.2 Web documents

The third main source of texts to be included in the micro-balanced corpus is the web. The web gives access to colossal quantities of texts of any type and more and more linguists and language technologists rely on it as a huge source of corpus materials (see Kilgarriff, 2001). The MEANING project itself treats the web as a corpus to learn information from it, with the final aim of opening the way for a concept-based access to the Multilingual Web.

Despite web's usefulness for corpus research, when trying to collect web documents we have to face several problems: the web contain duplicates or very similar documents, not all documents contain enough text, they may contain mixes of languages, and so on. As it is impossible to visit, download and manually classify some of the millions of web pages, we are at the moment studying how to devise automatic methods to draw materials from the web for inclusion in the corpus.

	Precision	Recall	Coverage
News	0.64	0.44	0.55

Table 4 *Performance of the second procedure on the news*

2.2 The macro-balanced component

The *macro-balanced* corpus is being created in order to evaluate in an independent way the relevance of the domains in a generic corpus. This corpus is not intended to be a reference corpus for the Italian language, as it is not balanced with respect to different literary genres, media, modes, and styles. It is a homogeneous corpus composed of two general high circulation newspapers ("La Repubblica" and "La Stampa") in which we expect most domains to be represented. The macro-balanced corpus contains about 90 million running words covering a time-span of 4 years (1998-2001). This time-span has been chosen in order to keep the corpus comparable with the other corpora of the MEANING consortium.

We assume that in the selected material the most common topics dealt with in periodical are represented, giving us a picture of the distribution and proportions of the topics within the corpus. This will allow us to verify the relevance of the different domains in the current language. Table 5 summarizes the data about the texts we included in the corpus.

	Size (tokens)	Time-span
La Repubblica	38 millions	2000-2001
La Stampa	48 millions	1998-1999

Table 5 *Structure of the macro-balanced component*

3. Corpus encoding

The corpus will be encoded using XML as a common data format. We will take into account, whenever possible according to the requirements of our NLP applications, the Corpus Encoding Standard for XML (XCES) guidelines and the new standard ISO/TC 37/SC 4 for language resources (Ide and Romary, 2002). We chose full text as type of sample for the corpus, that is the complete newspaper article, piece of news, or other document is taken as the minimum size of the text. Each text is stored in a separate file.

CES distinguishes three broad categories of information which are of direct relevance for the encoding of corpora for use in NLP applications:

- *Documentation*, which includes global information about the text, its content, and its encoding.
- *Primary data*, which consist of the text marked up with information regarding both gross structure (paragraphs, chapters, titles, footnotes, etc.; features of typography and layout; non-textual information, such as graphics, etc.) and sub-paragraph structures (sentences, highlighted words, dates, abbreviations, etc.)
- *Linguistic annotation*, i.e. information added to the primary data as a result of some linguistic analysis

In the MIC, documentation about each text will be included in the form of a separate XCES-conformant header. All the original texts are stored in the *legacy corpus*, which is kept as a backup corpus. Then, to obtain the encoded version of the corpus, the legacy texts undergo a series of transformations. To this extent, a number of normalization scripts have been implemented.

In the CES guidelines primary data (i.e. the text itself marked up with information about its structure) form the so-called *base* or *hub* text. The hub text does not include linguistic annotations which are stored in separate documents and linked to the hub text or other linguistic annotation documents. In the encoding of the MIC, we follow CES guidelines in retaining linguistic annotation in separate documents. However we differ from CES in the way we treat primary data. In fact, we prefer our hub corpus to be completely plain, i.e. pure text without any type of markup (apart from carriage returns). Thus the encoding of the primary data is not kept together with the text itself: primary level information is coded in the same way as linguistic information and is stored in different files separated from the hub text.

4. Corpus annotation

A multi-level annotation of the corpus is planned in order to encode seven different kinds of information: orthographic features, the structure of the text (primary data, level 1 and 2), morphosyntactic information, multiwords, syntactic information, named entities (primary data, level 3), and word senses.

All annotations are performed automatically, using linguistic tools developed at ITC-irst. Information about each level of annotation is stored in separate documents. Following the CES recommendations (see Section 3), all annotation documents are linked to the hub corpus or other annotation documents using one-way links. Two different means can be used to specify locations, namely reference to a unique identifier (ID) and reference to the position of the characters in the text We use the character position locators to link the orthographic annotation to the hub corpus. ID locators are used to link all the other linguistic annotation documents.

In the next sections, information about the different kinds of annotation are given. All the examples reported refer to parts of the same sentence: “Il Ministero della Sanità dice che coi superalcolici bisogna andarci veramente piano” (Eng. “The Department of Health and Human Services says that people must take it really easy with liquors”).

4.1 Orthographic annotation

The corpus is automatically tokenized and each token is annotated with:

- token ID
- Location in the hub corpus
- case (upper, lower, capitalized)

Example: “Il Ministero…” (The Department…)

```
<struct type="ortho">
 <struct type="t-level" id="t_1">
        <feat type="token">Il</feat>
        <feat type="case">capitalized</feat>
        <seg startsAt="0" endsAt="1"></seg>
 </struct>
 <struct type="t-level" id="t_2">
        <feat type="token">Ministero</feat>
        <feat type="case">capitalized</feat>
        <seg startsAt="3" endsAt="11"></seg>
 </struct>
 …
</struct>
```

4.2 Structure annotation

At this level of annotation, primary data (level 1 and 2) are encoded. As said before, this information is stored in a document separeted from the hub file, which contains only the pure text without any tags. The following information is recorded:

- text divisions, paragraphs, sentences, rendition information, etc.(i.e. all structural information)
- ID for text divisions, paragraphs, and sentences
- link to token IDs in the orthographic annotation file

Example: <p> Il Ministero della Sanità dice che coi superalcolici bisogna andarci veramente piano.
 Negli ultimi anni, infatti, il numero di cirrosi epatiche è in continuo aumento. <p> ...

```
<struct type="structure" xml:base="../ortho/ministero-ort.xml">

 <struct type="p-level" id="p_1"
      xlink:href="#xpointer(id('t_1')/range-to(id('t_29')))">
    <struct type="s-level" id="s_1"
        xlink:href="#xpointer(id('t_1')/range-to(id('t_13')))"></struct>
    <struct type="s-level" id="s_2"
xlink:href="#xpointer(id('t_14')/range-to(id('t_29')))"></struct>
 </struct>

 <struct type="p-level" id="p_2"
      xlink:href="#xpointer(id('t_30')/range-to(id('t_32')))">
    <struct type="s-level" id="s_3"
xlink:href="#xpointer(id('t_30')/range-to(id('t_...')))"></struct>
 </struct>

</struct>
```

4.3 Morphosyntactic annotation

After PoS tagging and lemmatization each token in the corpus is annotated with its morphosyntactic information, that is:

- word ID
- link to token ID in the orthographic annotation file
- lemma, stem, PoS, form (when necessary), morphological features (gender, number, mood, tense, person)

Moreover, if the word belongs to a multiword:

- link to the multiword ID in the multiwords annotation file
- function of the word in the multiword (head, satellite)

As regards POS tags, the tagset applied is a subset of the tagset specified in the EAGLES Guidelines for morphosyntactic annotation.

Example: “andarci veramente piano” (Eng. “take it really easy”)

```
<struc type="morpho" xml:base="../ortho/ministero-ort.xml">
…
<struc type="w-level" id="w_12" xlink:href="#xpointer(id('t_10'))">
  <feat type="lemma">andare</feat>
  <feat type="stem">and</feat>
  <feat type="form">andar</feat>
  <feat type="pos">v</feat>
  <feat type="elra-tag">VF</feat>
  <feat type="mood">inf</feat>
  <feat type="tense">pres</feat>
  <feat type="mwd-element"
```

```
       xlink:href="../multiwords/ministero-mwd.xml#xpointer(id('mwd_2'))">
       head</feat>
 </struc>

 <struc type="w-level"
       id="w_13"
       xlink:href="#xpointer(id('t_10'))">
  <feat type="lemma">ci</feat>
  <feat type="pos">pron</feat>
  <feat type="elra-tag">+E</feat>
  <feat type="mwd-element"
       xlink:href="../multiwords/ministero-mwd.xml#xpointer(id('mwd_2'))">
       satellite</feat>
 </struc>

 <struc type="w-level" id="w_14" xlink:href="#xpointer(id('t_11'))">
  <feat type="lemma">veramente</feat>
  <feat type="pos">avv</feat>
  <feat type="elra-tag">B</feat>
 </struc>

 <struc type="w-level" id="w_15" xlink:href="#xpointer(id('t_12'))">
  <feat type="lemma">piano</feat>
  <feat type="pos">avv</feat>
  <feat type="elra-tag">B</feat>
  <feat type="mwd-element"
       xlink:href="../multiwords/ministero-mwd.xml#xpointer(id('mwd_2'))">
       satellite</feat>
 </struc>

</struc>
```

4.4 Multiwords annotation

All expressions in the corpus which are multiwords are coded with the following information:

- multiword ID
- PoS, lemma
- link to the word ID of the components in the morphosyntactic annotation file
- function of the components words (head, satellite)

Example: “andarci piano” (Eng. take it easy)

```
<struct type="multiwords" xml:base="../morpho/ministero-mph.xml">
…
 <struct type="w-level" id="mwd_2">
     <feat type="lemma">andarci_piano</feat>
     <feat type="pos">v</feat>
     <struct type="mwd-element" xlink:href="#xpointer(id('w_12'))">
        <feat type="function">head</feat>
     </struct>
     <struct type="mwd-element" xlink:href="#xpointer(id('w_13'))">
        <feat type="function">satellite</feat>
     </struct>
     <struct type="mwd-element" xlink:href="#xpointer(id('w_15'))">
        <feat type="function">satellite</feat>
     </struct>
   </struct>

</struct>
```

If the multiword is present in our reference lexicon MultiWordNet[1] (Pianta et al. 2002), PoS and lemma are those of MultiWordNet.

4.5 Named Entities annotation

All named entities in the corpus are recognized and coded as such with the following information:

- named entity ID
- type of named entity
- link to the word ID or multiword ID in the respective annotation files

Example: Ministero della Sanità (Departement of Health and Human Services)

```
<struct type="namedentities" xml:base="../morpho/ministero-mph.xml">
 <struct type="ent-level" id="e_1"
```

[1] MultiWordNet is a multilingual lexical database, developed at ITC-irst, in which the Italian wordnet is strictly aligned with Princeton WordNet (version 1.6) (Fellbaum, 1998)

```
            xlink:href="../multiwords/ministero-
mwd.xml#xpointer(id('mwd_1'))">
  <feat type="enamex">organization</feat>
 </struct>
</struct>
```

The tagset applied to annotate named entities is the one adopted in the framework of the DARPA/NIST HUB4 evaluation exercise.

4.6 Word sense annotation

Content words and multiwords in the corpus which are present in MultiWordNet are disambiguated according to MultiWordNet synsets. The annotation includes:

- link to the word ID or multiword ID in the respective annotation files
- MultiWordNet lemma, PoS, and synset ID

Example: “bisogna andarci piano” (Eng. (people) should take it easy)

```
<struct type="semantic" xml:base="../morpho/ministero-mph.xml">

<struct type="sem-level"
        xlink:href="#xpointer(id('w_11'))">
  <feat type="MWN-lemma">bisognare</feat>
  <feat type="MWN-pos">v</feat>
  <feat type="MWN-sense">v#3990811</feat>
 </struct>

 <struct type="sem-level"
  xlink:href="../multiwords/ministero-mwd.xml#xpointer(id('mwd_2'))">
  <feat type="MWN-lemma">andarci_piano</feat>
  <feat type="MWN-pos">v</feat>
  <feat type="MWN-sense">v#03437782</feat>
 </struct>

</struct>
```

4.7 Syntactic annotation

Syntactic annotation will be carried out only in the last phase of the creation of the MIC. The precise encoding of the syntactic annotation has not been decided yet. However we plan to automatically annotate at least the main phrases of the sentence by using shallow parsing (phrase chunking) techniques.

5. Summary and conclusions

The MEANING Italian corpus has been presented in this paper. MIC is being developed in the framework of the MEANING project with the aim of supporting word sense disambiguation, however a domain-based corpus can be a very useful resource not only for natural language processing applications but also for different kinds of linguistic analyses.

The corpus is in its way to realization. All its overall structure has been designed and the multi-level annotation scheme has been developed. The macro-balanced component has been created, normalized and linguistically annotated up to level of morphosyntactic annotation. XCES-conformant headers for each texts have been automatically created. As regards the micro-balanced component, we are collecting materials from different sources and we are devising semi-automatic procedures to speed up its construction. Our work will go on until the corpus will be entirely created and all the levels of linguistic annotation will be performed.

Acknowledgments

We would like to thank Pamela Forner for her huge and precious work on the creation of the corpus and Claudio Giuliano, Nancy Ide, and Tomaz Erjavec who offered helpful comments for the development of the corpus annotation scheme.

References

Bentivogli L, Pianta E 2002 Opportunistic semantic tagging. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands - Spain.

Calzolari N, Baker M, Kruyt T (eds) 1995 Towards a network of european reference corpora. Report of the NERC Consortium Feasibility Study, coordinated by Antonio Zampolli. *Linguistica Computazionale* XI-XII. Pisa, Giardini.

Dewey Decimal Classification and Relative Index 1996, Ed. 21, edited by J.S. Mitchell, Forest Press, Albany.

Fellbaum C (eds) 1998 *WordNet: An Electronic Lexical Database*. Cambridge(Mass.), The MIT Press.

Grishman R, Kittredge R (eds) 1986 *Analyzing Language in Restricted Domains*. Lawrence Erlbaum.

Ide N, Romary L 2002 Standards for Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands – Spain.

Ide N, Macleod C 2001 The American National Corpus: A Standardized Resource of American English. In *Proceedings of Corpus Linguistics*, Lancaster, UK.

Kilgarriff A 2001 Web as corpus. In *Proceedings of Corpus Linguistics 2001 Conference*, Lancaster,UK.

Magnini B, Strapparava C, Pezzulo G, Ghiozzo A 2003 The Role of Domain Information in Word Sense Disambiguation. *Journal of Natural Language Engineering* (special issue on Senseval-2) 9(1).

Magnini B, Cavaglià G 2000 Integrating Subject Field Codes into WordNet. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

Magnini B, Gliozzo A 2002 *Mapping WordNet Domains to the Dewey Decimal Classification*. ITC-irst Technical Report.

Montemagni S, Barsotti F, Calzolai N, Corazzari O, Zampolli A, Fanciulli F, Massetani M, Raffaelli R, Basili R, Pazienza M T, Saracino D, Zanzotto F, Mana N, Pianesi F, Del Monte R 2000 Building the Italian Syntactic-Semantic Treebank. In *Building and using syntactically annotated corpora*. Kluwer, Dortrecht.

Pianta E, Bentivogli L, Girardi C 2002 MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First Global WordNet Conference*, Mysore, India.

Rigau G, Magnini B, Agirre E, Vossen P, Carroll J 2002 MEANING: A Roadmap to Knowledge Technologies. In *Proceedings of COLING Workshop "A Roadmap for Computational Linguistics"*. Taipei, Taiwan.

Rossini Favretti R, Tamburini F, De Sanctis C 2001 A corpus of written Italian: a defined and dynamic model. In *Proceedings of Corpus Linguistics 2001 Conference*, Lancaster,UK.

Sekine S 1997 The Domain Dependence of Parsing In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington D.C., USA.

Sinclair J, Ball J 1996 *EAGLES Preliminary Recommendations onText Typology*. (http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html)