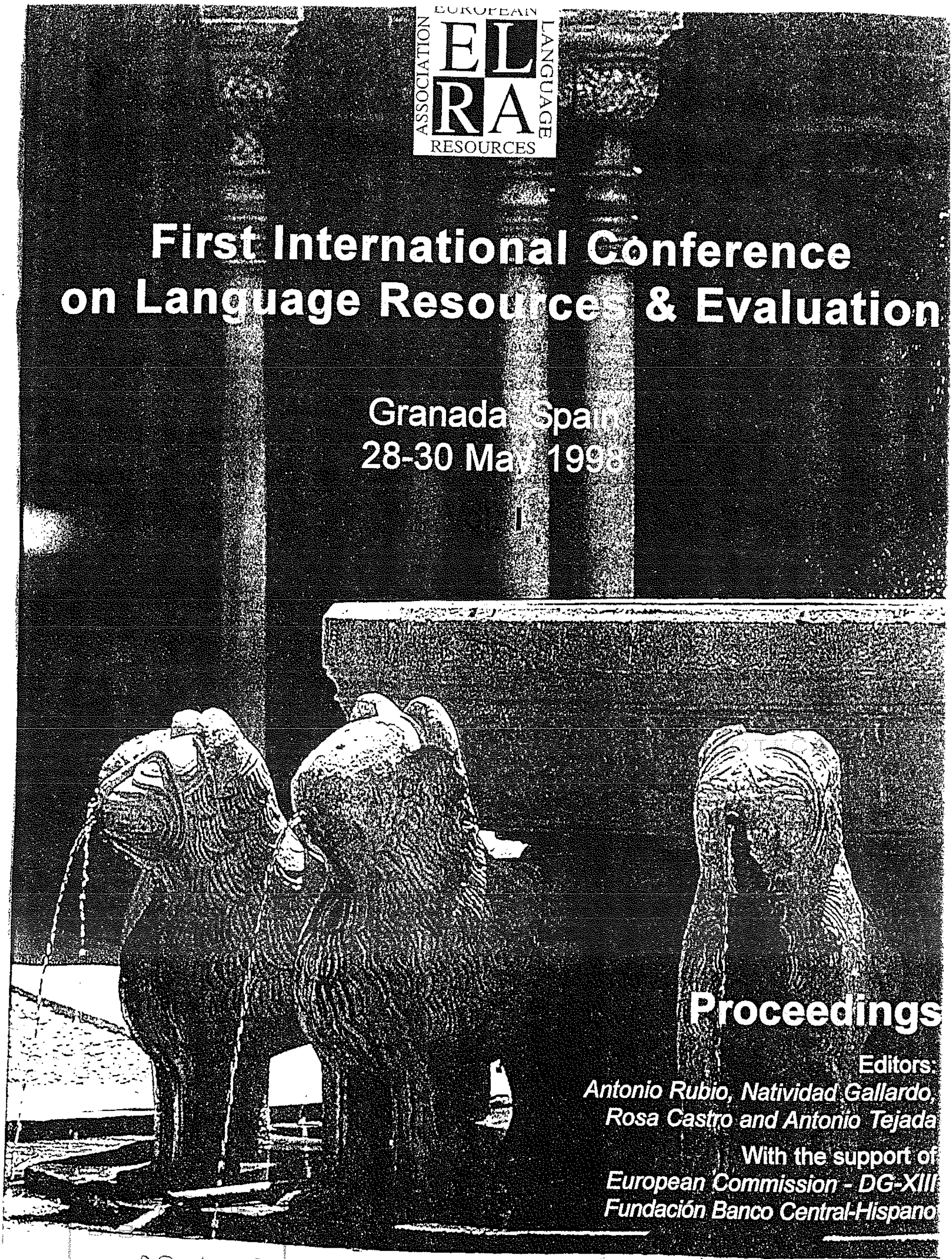
EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

First International Conference on Language Resources & Evaluation

Granada, Spain
28-30 May 1998

I



Proceedings

Editors:
Antonio Rubio, Natividad Gallardo, Rosa Castro and Antonio Tejada

With the support of
European Commission - DG-XIII
Fundación Banco Central-Hispano

European Language Resources Association

First International Conference on Language Resources and Evaluation

Title: First International Conference on Language Resources and Evaluation
Volumes: I & II
Editors: Antonio Rubio, Natividad Gallardo, Rosa Castro y Antonio Tejada

Additional copies may be ordered from:
ELRA - European Language Resources Association
55-57 Rue Brillat-Savarin, 75013 , Paris
France
Tel. +33 1 43 13 33 33
Fax. +33 1 43 13 33 30

Printed in:
Copisteria "La Gioconda"
Melchor Almagro, 16
Granada (Spain)
D. L. GR: 623-98

Cover page design: Rébecca Jaffrain

Photograph front page: "Andalousie : l'Alhambra", *Photographer*: Borredon T.
Photograph back page: "Espagne, Andalousie, Grenade, La Cour des lions", *Photographer*: Sappa C.

SESSION M: LEXICAL PROJECTS (2): SEMANTICS NETS

SESSION N: EVALUATION: TOKENIZERS, TAGGERS, PARSERS

SESSION O: CORPUS PROJECTS

SESSION P: SPOKEN LANGUAGE RESOURCES PROJECTS (2)

SESSION Q: LANGUAGE RESOURCES: STRATEGIC ISSUES

KEYNOTES

SESSION R: ONTOLOGIES & KNOWLEDGE BASES

Building a Semantic Network for Italian using Existing Lexical Resources

Adriana Roventini[1], Carol Peters[2], Nicoletta Calzolari[1], Francesca Bertagna[1]

[1] Istituto di Linguistica Computazionale, CNR, Pisa
[2] Istituto di Elaborazione della Informazione, CNR, Pisa

Abstract

The paper describes the approach adopted to construct the Italian component of the EuroWordNet multilingual semantic database. A significant part of our work has been to exploit as far as possible already existing lexical resources and data processing tools and to use the results of previous projects in order to optimize the work necessary to build the monolingual database. We will discuss the main problems encountered when restructuring our source data as a semantic net in which word meanings are represented by groups of similar word senses (the "synsets") and the web of semantic relations holding in the lexicon between the different word meanings is rendered explicit. The strategies adopted to ensure that the Italian net is linguistically coherent while guaranteeing compatibility with the databases for the other languages are also discussed and the results are presented.

Introduction

The aim of EuroWordNet[1] is to construct a multilingual semantic database in which different monolingual wordnets for European languages (in the first phase Dutch, (British)English, Spanish and Italian) are linked through an Interlingual Index or ILI (EWN 96-97), which is essentially a modified version of WordNet 1.5 (Miller et al, 1990). In the paper, we describe how the Italian component of EuroWordNet is being built, mainly from existing resources, and the strategies adopted to ensure that it is linguistically coherent while guaranteeing compatibility with the databases for the other languages. (For a full description of the EuroWordNet project - its objectives, development and results - see the special edition of Computers and the Humanities, to appear shortly).

As is known, the original WordNet was built from scratch at the Cognitive Science Laboratory of Princeton University on the basis of psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. We have taken the WordNet model as the starting point in the construction of our Italian lexical/semantic database, and our core relation is based on the Wordnet "synset".

Like WordNet, we have adopted a somewhat weak definition of synonymy, i.e. semantic similarity, entailing the interchangeability of words in a particular context. But, differently to the Princeton group, we have extended this idea to cover different syntactic categories, taking the idea of interchangeability to refer to a deeper semantic level rather than the superficial syntactic realization. We have thus increased the range of the different types of relations which are encoded, in particular to include cross part-of-speech semantic connections. However, it is the way in which we are building the Italian WordNet that really differentiates our work from that of Princeton. Our objective has been to investigate to what extent it is possible to construct a large, complex semantic net by restructuring already existing lexical resources and adapting the methodologies, techniques, and tools that had been developed to process them.

In adopting this approach, we have had two main motivations: to economize on the effort needed to build a database of this type by exploiting the results of previous projects; to achieve an objective perspective by abstracting away from the idiosyncrasies of a single resource or a particular theory, through the merging of data from a number of different mono- and bilingual electronic dictionaries and lexical databases.

Our goal has been: (i) to construct a flexible tool for certain Italian NLP tasks requiring a semantic component, such as sense disambiguation; (ii) to create a component for a multilingual database (EuroWordNet) that can be used in cross-language studies and multilingual information retrieval.

The paper is organised as follows. In the next section, we illustrate how we have built the Italian WordNet from existing lexical resources available at the Istituto di Linguistica Computazionale (ILC-CNR), Pisa, the benefits we have obtained by using previously analysed data, and the problems that we have had to address in structuring the database and encoding the semantic relations. The following section describes how the monolingual database has been linked to the EuroWordNet Interlingual Index through a set of cross-language equivalence relations. In the final section, we describe our results quantitatively in terms of the numbers of entries and associated semantic relations, and comment them. Our aim is to show how we have been able to construct a linguistically coherent wordnet mainly by restructuring and building on existing resources, thus reducing the effort required.

[1] EuroWordNet (LE4003) is a project in the EC Language Engineering programme. The project partners are: University of Amsterdam (coordinator), Fundacíon Universidad Empresa (a cooperation of UNED Madrid, Politecnica de Catalunya, Barcelona, and the University of Barcelona), University of Sheffield, Istituto di Linguistica Computazionale, CNR, Pisa, and Novell Linguistic Development (Antwerp). In a second stage, the EWN database will be extended with French, German, Czech, and Estonian.

The Italian WordNet

When constructing the Italian wordnet, our primary commitment was to provide an adequate representation of the particular features of the Italian lexical system. In order to do this, we decided to merge data from a number of different sources. In this way, we felt that we could ensure a better and more accurate coverage of our language: the pecularities and idiosyncrasies of a single source could be eliminated by a comparison with other archives. On the other hand, important distinctive features of the language were evidenced by the fact that they received similar treatment in all of the sources, indifferently.

The first stage of our construction procedure was as follows: (i) we defined an initial core vocabulary for Italian, (ii) we encoded the main language internal semantic relations for this subset, and only then (iii) we mapped our data entries to the project Interlingual Index (ILI). In a second stage, after a comparison with the first results of our partners (matching over the ILI), we revised and integrated this core subset and began to expand it downwards to the first levels of hyponyms.

Existing Resources

In building our WordNet, the lexical resources, the tools and the methodologies that have been developed at ILC-CNR in recent years played an essential role. In particular, we were able to exploit the results of previous EC language engineering projects (e.g. Acquilex ESPRIT BRA and LRE-Delis) in which dictionary entries and definitions had been analysed and many kinds of semantic relations had been identified - starting from machine readable dictionaries and from corpora - and (partially or completely) encoded. This information had been added to our main source, the Italian Lexical Database (LDB).

The LDB is a very large lexical data archive built up over the years on the basis of data from a number of different sources. The subset of this archive used for the Italian WordNet currently consists of nearly 25,000 of the most central entries (approximately 20,000 nouns and 5,000 verbs). This number will be increased by the end of the project to about 35,000 entries. The entries in the LDB have all been analysed and the information extracted has been tagged. The following semantic relations had already been partially encoded: synonymy, hyperonymy/hyponymy, part-of, set-of, agent of, deverbal, deadjectival for nouns; synonymy, hyponymy/hyperonymy and causative/inchoative alternation for verbs. Table 1 shows a few examples of lexical entries from the LDB in which different kinds of semantic relations have been encoded; the semantic relations that are extracted for the Italian Wordnet are given in bold.

Has_Hyperonym relations (isa=hyperonymy)
AFFLIZIONE {lemma} *afflizione* {senso} 0-0 {defin} "stato di grande abbattimento." {relaz} isa {conte} **stato** {sencn} 0-2

DOLORE {lemma} *dolore* {senso} 0-0 {defin} "sensazione di sofferenza, di molestia, di pena, causata da un male fisico o morale." {relaz} **isa** {conte} **sensazione** {sencn} 0-1

Has_Holo_Member relation (isacoll=set of)
ARGENTERIA {lemma} *argenteria* {senso} 0-0 {defin} "insieme di oggetti d'argento, spec. stoviglie e vasellame." {relaz} **isacoll** {conte} **insieme** {relaz} **setof** {conte} **oggetto**

Has_Mero_Part relation (isapart=part of)
PARAFANGO {lemma} *parafango* {senso} 0-0 {defin} "in un veicolo, parte della carrozzeria che copre parzialmente le ruote ." {relaz}
isapart
{conte} **parte** {sencn} 0-1 {relaz} **partof** {conte} **carrozzeria**

Be_In_State relations (a2n=deadjectival)
GENTILEZZA {lemma} *gentilezza* {senso} 0-0a {defin} "l'essere gentile." {relaz} **a2n** {conte} **gentile** {deriv} deriv {relaz} isaa2n {conte} essere {senso} 0-0b {defin} "atti o parole gentili." {relaz} isa {conte} atto {sencn} 1-1 {relaz} isa {conte} parola

CORTESIA {lemma} *cortesia* {senso} 0-1 {defin} "l'essere cortese." {relaz} **a2n** {conte} **cortese** {deriv} deriv {relaz} isaa2n {conte} essere {senso} 0-2 {defin} "atto cortese." {relaz} isa {conte} atto {sencn} 1-1

XPOS_Near_Synonym relations (v2n=deverbal)
ABBELLIMENTO {lemma} *abbellimento* {senso} 0-1a {defin} "atto dell'abbellire." {relaz} isav2n {conte} atto {relaz} **v2n** {conte} **abbellire** {deriv} deriv {senso} 0-1b {defin} "ornamento." {relaz} syn {conte} ornamento {sencn} 0-0b {senso} 0-2 {defin} "note

ABBATTIMENTO {lemma} *abbattimento* {senso} 0-1a {defin} "atto, effetto dell'abbattere." {relaz} isav2n {conte} atto {relaz} isav2n {conte} effetto {relaz} **v2n** {conte} **abbattere** {deriv} deriv {senso} 0-1b {defin} "demolizione." {relaz} syn {conte} demolizione {sencn} 0-0

Role_Agent relations (agof=agent of)
LETTORE {lemma} *lettore* {senso} 0-1a {defin} "chi legge." {relaz} **agof** {conte} **leggere** {relaz} tgt {conte} chi {senso} 0-1b {defin} "chi legge con particolare intensit...." {relaz} agof {conte} leggere {relaz} tgt {conte} chi {senso} 0-2 {defin} "chi ha ricevuto il secondo

ACCORDATORE {lemma} *accordatore* {senso} 0-0 {defin} "chi accorda strumenti musicali, spec. pianoforti." {relaz} **agof** {conte} **accordare** {relaz} tgt {conte} chi

Table 1: Examples of how Semantic Relations had been previously encoded in the Italian Lexical Database

Our other main data sources were the following:

- an electronic synonym dictionary, used as a source of indications on synonym data and word-sense distinctions;
- an Italian/English bilingual LDB, containing approximately 30,000 entries in each dataset. It is used to provide a first translation of Italian word-senses (see following section), and also as a source of potential synonyms, providing a different perspective from that of the monolingual LDBs;
- the Italian Reference Corpus, used as an additional source of data, and as confirmation of current usage.

Building the Italian WordNet

As in WordNet, the primary relation in the monolingual database is the synset. Our first step was thus to build a core set of synsets representing what should be the base concepts in the Italian lexical system. This was in accordance with one of the main decisions of the EuroWordNet project: a vocabulary subset should be selected for each participating language. This subset should represent the most general and commonly used word-senses in that language (the criterion adopted was "word-senses most frequently used to define other words in dictionaries") in such a way that (i) no important lexical/semantic area was neglected, (ii) the highest taxonomic levels for the entire lexicon were covered. In this way a high degree of compatibility between the separate monolingual databases was ensured. The selection of this first set of language-dependent "base concepts", easily extracted by means of the disambiguated is-a relations, was followed by a stage of cross-language comparison in order to be able to establish a common set of word-senses representing these "base concepts" for all the four EuroWordNet languages. In order to be able to do this, all the base concepts for each project language were linked manually to the InterLingual Index (at this stage entirely represented by WordNet 1.5).

A second decision of the project was to give even more attention than WordNet to the idea that the lexicon is a network of relations. A number of additional semantic relations were thus included, in particular cross part-of-speech relations, e.g. from noun to verb, or noun to adjective, etc. This decision meant that basic word-meanings are represented from many perspectives and in many contextualisations. The aim of the component wordnets of EuroWordNet is to provide a lexical representation of a given language rather than a conceptual representation; however, by encoding cross part-of-speech semantic relations, the semantic structure of EuroWordNet is perhaps represented at a deeper level than that of WordNet. This decision was motivated by the consideration of a number of potential applications of EWN, such as information retrieval and machine translation, where it is essential to have a link between different lexicalizations (also with different parts of speech) having the same underlying meaning.

The major problems we had to face in developing the database on the basis of existing source data regarded the typical incoherencies found in lexicographic metalanguage such as circularity, under and over sense differentiation, inconsistency in the "genus" or hyperonym assignment, hyperonym disjunction or conjunction. For instance we decided that disjunction had to be eliminated as a cause of ambiguity. This implied splitting such entries into distinct meanings corresponding to different hyperonyms and, frequently, separate taxonomies. This decision often has repercussions on the hyponyms of the divided entries as we move down the taxonomies: for example, when we are processing meanings that are strictly related, such as those entries originally defined in our sources as *atto* or *effetto* (act or effect), it is not enough to simply create two separate entries at the base concept level, but each hyponym must be afterwards evaluated to establish whether it inherits both meanings or just one.

Creating Synsets

The construction of the first set of synsets for the Italian Wordnet and its subsequent extension was performed semi-automatically.

The first step was to automatically extract the set of base concepts from our LDB, using the criteria stated above. Those entries which were sufficiently high in our taxonomies (taxonomic chains had already been constructed, in the ACQUILEX project, on the basis of the hyponym/hyperonym links) and had the highest number of hyponyms were selected. This set was revised manually and obvious gaps were filled, e.g. sets of family relationships were completed, as were sets of entries referring to atmospheric phenomena, etc.. This set provides the basic layer of more generic word-senses, from which all the other specific senses can be derived through various levels of is-a links.

We then built our first synsets for each of our base concept word-senses, i.e. looked for their synonyms. An automatic procedure associated with each "base concept" all word-senses that (i) were explicitly tagged as synonyms in the LDB, (ii) were provided with synonymical type definitions (see the example below: *amore*, and both senses of *devozione* have been given synonym type definitions), or (iii) were listed as synonyms in the electronic synonym dictionary. Each word-sense was also searched in the bilingual dictionary as frequently a synonym of the sense is provided as a semantic indicator before a translation. The results of this automatic stage had to be then checked and revised manually. Frequently in fact, the synsets formed in this way for the base concepts were too large and had to be restricted by cutting irrelevant senses. However, sometimes the group of synonyms proposed by the automatic procedure was found to be well-formed without manual intervention, as can be seen for the example of *affetto* (affection) taken from the taxonomy for {feeling, emotion}. A completely automatic search gave us the following chain:

affetto 0_1 SYN *affezione*, *amore*, *tenerezza*
affezione 0_1b SYN *affetto*, *attaccamento*
amore 0_1c DEFSYN *devozione*
attaccamento 0_0 SYN *affezione*
tenerezza 0_2 (fig) SYN *affetto*
devozione 0_2a DEFSYN *dedizione* SYN *affezione*, *fedeltá.*
devozione 0_2b DEFSYN *attaccamento* SYN *affezione*, *fedeltá*
dedizione 0_0
fedeltá 0_1 SYN *devozione*

At this point, by gathering the words explicitly indicated as synonyms *affetto*, *affezione*, *amore*, *tenerezza*, *attaccamento*, and the synonyms by definition type (DEFSYN) *devozione* and *dedizione* we obtain the synset:

{*affetto* 0_1 (affection), *affezione* 0_1b (fondness), *amore* 0_1c (love), *attaccamento* 0_0 (attachment), *tenerezza* 0_2 (tenderness), *devozione* 0_2a (devotion) *dedizione* 0_0 (devotion)*fedeltá* 0_1 (faithfulness)}.

This could probably be enlarged or restricted depending on the linguistic competence of each speaker, but is, in any case, an acceptable result. In fact, if we

look at the WordNet 1.5 data, we find this synset for affection:
{affection, affectionateness, fondness, tenderness, heart, warmheartedness}, which is very similar to the Italian one.

Deepening Taxonomies and Encoding Relations

Once we had successfully created the set of synsets for our base concepts, we began to construct the taxonomies, top-down, using the taxonomic chains that had already been automatically built for the LDB data, and manually checking, revising and extending them. During this stage, a serious problem was evidenced by the analysis of the hyponyms: the lack of depth in our taxonomies which, in some cases, presented an enormous heterogenity in the elements that had been grouped together at the same level by too generic defining formulae. The most obvious case of this is provided by the set of Agents which, in our source database, are mainly defined with the formula *chi* (pronoun) + VP (*chi* = who has been mapped to the base concept synset {*persona*, *individuo*, *essere_umano*, *uomo*} {person, individual, human_being, man}, therefore hyponyms of *chi* are considered as hyponyms of *person*). This means that 2000 items have been automatically tagged as agents. However, checking and evaluating this set of data we found that, in order to render it significant and well structured, a considerable manual intervention was needed to distinguish e.g. the various specific types of agents: *operaio*, *artigiano*, *artista*, *venditore*, *fabbricante*, *negoziante* (types of human occupation) from very generic agents such as *comunicatore*, *ammiratore*, *calunniatore*, etc. (actions that one can perform) and also from patients of the type *chi è affetto da*, *chi soffre di*, etc. (who is affected by, who suffers from, etc.). This redistribution and regrouping of hyponyms at different levels of specificity within the taxonomy - and therefore addition of intermediate nodes - implied considerable manual work. Moreover, for each entry that we added to the Italian WordNet in construction, we also extracted any semantic relations that had been explicitly encoded in the LDB entries. For nouns, apart from the hyperonym/hyponym and synonym relations, the main relations, for which we had a large and fairly consistent existing set of data were cross part-of-speech relations (of derivational type), agent and some mero-holonym relations. Other relations were encoded by hand. For example, we can cite the case of the term *edificio* (building) and its hyponyms. This term automatically selects 82 hyponyms. Analysing their definitions and inserting manually those semantic relations which could not be automatically extracted from the "differentia" part of the definitions, we had to add to our data a great number of relations such as Role-location, Has_mero_part, Has_holonym. The definitions in this subset generally evidence the role of the building, but in some cases many structural elements are also listed. Encoding all of them implied that manual work increased considerably. Let us give just one example: *castello* (castle) is defined as: *grande edificio munito di mura e torri, circondato da un fossato, in cui abitavano i signori feudali* (large building fortified with walls and towers, surrounded by a moat, where feudal lords lived). We can thus derive:

Role_location: *abitare* (to live in)
Has_mero_part: *mura* (walls)
Has_mero_part: *torre* (tower)
Has_mero_part: *fossato* (moat)

This means that we have manually inserted four internal relations in addition to the hyperonym relation which had been obtained automatically. And *castello* is not an isolated case; in Figure 1 (found at the end of the text) shows the base concept synset *nave*, *bastimento* (ship) with hyponyms and other encoded semantic relations as it appears in the EWN tool viewer; Table 2 below shows these relations from a quantitative point of view: the asterisk indicates those relations which were extracted (almost entirely) automatically.

Language Internal Relations	Nouns
*Be_In_State	123
*Has_Hyperonym	18654
Has_Holo_Location/Member/Made_of	41
*Has_Holo_ Part	290
*Has_Meronym	264
Has_Mero_Madeof	165
Has_Mero_Member	186
*Has_Mero_Part	219
Has_Mero_ Location	5
*Near_Antonym	20
Near_Synonym	221
Role	21
*Role_Agent	1095
Role_Instrument	80
Role_Location	51
Role_Patient	16
*XPOS_Near_Synonym	7505

Table 2: Language Internal Relations for Nouns

Mapping to the Interlingual Index (ILI)

In EuroWordNet, all the language-specific nets are stored in a unique database. Equivalence relations between entries in different languages are made explicit via the ILI. This is an unstructured version of WN1.5, in which original senses are be modified and new senses added when necessary. Each synset in the monolingual wordnet have at least one equivalence relation with an ILI record which enables cross-language mapping and comparison. This can be an equivalent synonym relation when there is an exact matching between the Italian and the English data, an equivalent near-synonym relation when the match is close but not precise, and an equivalent hyperonym relation when we are dealing with language specific objects that have no match in the other language. Linked to the ILI - on top of it - is a language independent Top Ontology and a set of domain labels.

We have deloped a semi-automatic procedure to establish the equivalence relations between the Italian data and the Wordnet synsets. We attempt to match the lexical semantic taxonomies that we had constructed for the Italian database against equivalent taxonomies in WordNet 1.5; it is the semantic context provided by the taxonomies that allows us to recognise the right sense in the target language of the word we are examining. Thus, although the ILI itself is unstructured, we have exploited the structure of WordNet 1.5 in order to make the right connections between the Italian lexical entries and the WordNet senses.

Our mapping procedure operates taxonomy by taxonomy. We start with base concepts that have already been mapped manually to our ILI through WordNet 1.5 and therefore provide us with a set of accurate anchor points between the Italian database and WordNet 1.5. Then, working top-down, we take all the first level hyponyms for each Italian base concept and input them to our bilingual lexical database system. For each word, all possible translations are read; we then search in the equivalent semantic hierarchy in WN1.5 - identified using the base concept links - in order to find an entry that matches one of the candidate translations; the assumption is that matching entries in equivalent semantic herarchies in different languages will refer to equivalent senses.

For example, in the taxonomy for Insects, the bilingual LDB assigns three possible translations to the Italian form *farfalla*: butterfly, bow-tie, butterfly-stroke. In the WN1.5 hierarchy under Insect, only one of these translation candidates was found (butterfly) and a link was thus created.

However, mapping is not always so straightforward. When no WN1.5 equivalent is identified, the procedure maps the Italian entry with a Has_equivalent_hyperonym relation to the WN1.5 base concept taken as the anchor point. The results of the first stage of the mapping procedure thus have to be checked and integrated manually. The following factors considerably affect the performance of the procedure:

- Our bilingual LDB is small - approximately 30,000 Italian-English entries and thus much translation data is missing.
- Differences in lexicalization and cultural differences limit the possible number of exact equivalences over languages.
- There may be no equivalent in WN1.5 to the Italian entry.
- There may be more than one possible WN1.5 equivalent in the same taxonomy.

All entries that are not assigned an equivalent_synonym relation automatically by the procedure are encoded manually. It is not always easy to recognize equivalences when they are not provided specifically by the bilingual dictionary (e.g. very domain-specific knowledge is required to identify the translations of many plants, insects, measuring instruments, etc.). The main difficulties we encountered were differences in lexicalization, mismatches and cultural gaps:

1. Differences in Lexicalization

Italian uses gender far more than English. Thus, in our taxonomy for *donna* (woman) we have many entries such as *ladra* (woman thief), *avvocatessa* (woman lawyer), etc. These are mapped as Eq_near_synonyms to the WN1.5 gender-neutral equivalents: thief, lawyer, etc. The information that this particular thief or lawyer is feminine is thus lost. Another major difference between a Romance language such as Italian and Germanic languages such as Dutch or English is in the use of multiwords; the Germanic languages are far richer in recognized multiwords and compounds. This means that lexicalized equivalents do not always exist, e.g. English toenail cannot be matched directly to the Italian equivalent "unghia del piede" not because the concept does not exist but because it is not lexicalized in the same way.

2. Mismatches and Lexical Gaps

When it is not possible to establish a direct equivalent synonym relation between our data and an ILI record, other kinds of equivalence relations are used, For example, Italian makes a clear distinction between hair-on-the-head *capelli* and hair-on-the-body *peli*. Both these word senses will be mapped to the ILI record for "hair" with an "Has_equivalent_hyperonym" relation. On the contrary, relations of equivalent hyponymy will be established between Italian *dito* and the ILI records for "finger" and "toe". The majority of lexical gaps are caused by cultural differences. As is to be expected, this is very evident in the Food taxonomy. Many cases are solved either by "Has_equivalent_near_synonym" relations (e.g. between *polpetta* and meatballs or rissoles) and "Has_equivalent_near_hyperonym" relations (e.g. between *castagnaccio* and cake).

Benefits

In any case, the procedure helps us to speed up the mapping process and permits the data to be treated more exhaustively and with more overall accuracy. As the data is organised taxonomy by taxonomy, comparison over languages is facilitated and candidate equivalences are more easily identified.

Additional advantages of using the procedure are:

1. Verification of our Taxonomies.

As described, our taxonomies have been built automatically from our dictionary source data; it is thus not surprising that they contain many imprecisions. By mapping them against the WN1.5 semantic taxonomies, which have been built carefully from scratch, differences are clearly evidenced. When the matching WN1.5 is found in a different taxonomy (during the manual evaluation stage), the Italian data is re-evaluated. For example, our taxonomy for *abitante* (inhabitant) - under *persona* (person) - includes *marziano*, *venusiano*, *selenita*. The first two are defined as hypothetical inhabitants of the planets Mars and Venus, respectively; the third as inhabitant of the Moon (without even the indication of hypothetical). WN1.5 contains only an equivalent for Martian and this appears in a very different taxonomy: Imaginary Being which lies under Psychological Feature. We have to decide whether to keep the Italian entries under the *persona* taxonomy or reclassify them, in this case introducing a new taxonomy into the Italian net.

2. Addition of Structure

One major problem when building the Italian WordNet was the tendency of our automatically derived taxonomies to be too shallow, especially with respect to the WN1.5 hierarchies. As has been stated, in many cases, it was necessary to restructure them, adding intermediate levels for too large sets of hyponyms in which many very specific terms were directly linked to hyperonyms that were too high and too generic. A comparison of our results with similar data in WN1.5 evidences obvious gaps in our lexical entries and also shows where some necessary structure can be added to our too-shallow hierarchies, e.g. following the example of WN1.5 but still respecting our data and on the basis of information provided in our LDB definitions, we have subdivided our taxonomy for "mammals" into ruminants, rodents, aquatic mammals, etc. Similarly, in the "instruments" taxonomy we introduced multiwords, which do not appear as lexical entries in the Italian monolingual LDB but are recognised lexicalised expressions such as *strumenti musicali* (musical instruments), *strumenti di misura* (measuring instruments), *strumenti di bordo* (navigation instruments), to create a new level in the taxonomy and, at the same time, to identify more homogeneous lexical subsets.

3. Automatic construction of synsets

Another important benefit of the mapping procedure is that it allows us to group similar Italian word-senses automatically, as the procedure moves downwards through the taxonomies. When two or more Italian entries have been linked to the same WN.5 word-sense, then an Italian synset is created with no need for manual intervention. The cross-language mapping procedure is thus proving effective not only in linking our Italian WordNet to the EWN Interlingual Index but also by providing valuable feedback, evidencing interesting lexical gaps and helping us to improve the coherency and consistency of our monolingual database.

Evaluation of Results

In our opinion, there were two main advantages in constructing our Italian WordNet from already existing and previously analysed lexical resourses:

- we had a large quantity of explicitly tagged lexical data in machine readable form, structured in a way that it could be easily reformatted - this meant a huge saving in the costs of the initial preparation of the data;
- the lexical data in our archives had orignally been extracted from a number of different authoritative sources and extended by the addition of further information taken from the Italian Reference Corpus, e.g. new terminology, information on usage, etc.. The analysed data is a result of different studies at ILC-CNR by lexicographers and linguists over the last years. It can thus be taken as a reliable representation of the Italian lexical system.

Here below we give some figures that give an idea of the resources that have been needed to construct the core dataset of the Italian net. These figures refer to the data delivered at the 24 month project milestone and to an effort for the Italian group of 26 man months.

	Nouns	Verbs	Total
Synsets	18934	3692	22626
No. of word-senses	19646	4588	24234
Senses per synset	1.03	1.24	
Language Internal Rels	47090	9070	
Equivalent Rels to ILI	7350	2090	
Average per Synset	0.27	0.17	

Table 3: Overall Figures for the Italian WordNet core data set

Equivalence Relations (to the ILI)	Nouns
Has_Eq_Synonym	4772
Has_Eq_Near_Synonym	631
Has_Eq_Hyperonym	1947
Total	7350

Table 4: Cross-Language Equivalence Relations for Nouns

References

EWN96-97: EuroWordNet documentation and papers, Http://www.let.uva.nl/~ewn/docs.htm.

Alonge A., Calzolari N., Hagman J., Marinai E., Montemagni S., Picchi E., Peters C., Roventini A., Spanu A., Zampolli A., Copestake A., Sanfilippo A., Vossen P. (1991). "An Overview of Work on Semantic Taxonomies in Pisa". ACQUILEX, ESPRIT BRA 3030, April 1991.

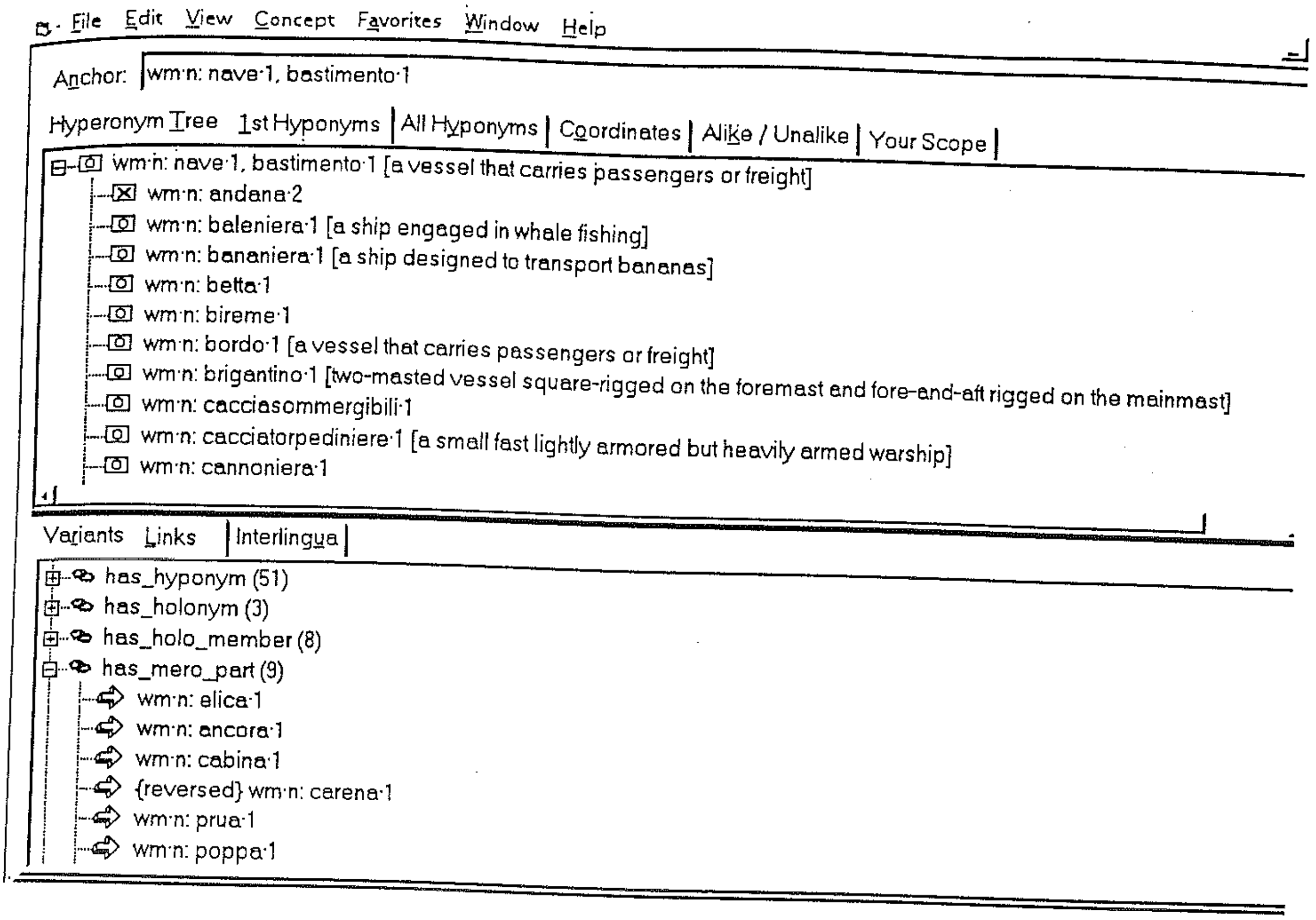
Climent S., Rodríguez H., Gonzalo J. (1996). "Definition of the links and subsets for nouns of the EuroWordNet project", EuroWordNet Project LE4003, Deliverable D005, October 1996.

Miller G.A, Beckwidth R., Fellbaum C., Gross D., and Miller K.J. (1990). "Introduction to WordNet: An On-line Lexical Database", in: International Journal of Lexicography, Vol 3, No.4 (1990), 235-244.

Miller G.A. (1990). "Nouns in WordNet: a Lexical Inheritance System", in: International Journal of Lexicography, Vol 3, No.4 (1990), 245-263.

Monachini M., Roventini A., Alonge A., Calzolari N., Corazzari O. (1994). "Linguistic Analysis of Italian Perception and Speech Act Verbs, Istituto di Linguistica Computazionale", Pisa, LRE-DELIS, Final Report, February 1994.

Vossen, P. (1997). "EuroWordNet: a multilingual Database for Information Retrieval", in Third DELOS Workshop "Cross-Language Information Retrieval", Zurich, 5-7 March 1997, ERCIM-97-W003, pp 85-93.



Anchor: wm·n: nave·1, bastimento·1

- wm·n: nave·1, bastimento·1 [a vessel that carries passengers or freight]
  - ☒ wm·n: andana·2
  - wm·n: baleniera·1 [a ship engaged in whale fishing]
  - wm·n: bananiera·1 [a ship designed to transport bananas]
  - wm·n: betta·1
  - wm·n: bireme·1
  - wm·n: bordo·1 [a vessel that carries passengers or freight]
  - wm·n: brigantino·1 [two-masted vessel square-rigged on the foremast and fore-and-aft rigged on the mainmast]
  - wm·n: cacciasommergibili·1
  - wm·n: cacciatorpediniere·1 [a small fast lightly armored but heavily armed warship]
  - wm·n: cannoniera·1

- has_hyponym (51)
- has_holonym (3)
- has_holo_member (8)
- has_mero_part (9)
  - wm·n: elica·1
  - wm·n: ancora·1
  - wm·n: cabina·1
  - {reversed} wm·n: carena·1
  - wm·n: prua·1
  - wm·n: poppa·1

Figure 1: Semantic and Equivalence Relations encoded for {nave, bastimento}