Acknowledgements

We would like to thank all those who, through their intellectual involvement, their logistic support or financial assistance, made this Eighth Euralex International Congress possible. We are particularly grateful to the following public or private organisations:

Cambridge University Press
Cercle Belgo-Néerlandais, Université de Liège
DGXIII/E5 (European Commission, Telecommunications, Information Market and Exploitation of Research : Language Engineering and Applications)
Euralex Executive Board
Faculté de Philosophie et Lettres, Université de Liège
The Hornby Trust
Interface Entreprises-Université, Liège

Ce congrès a également bénéficié du soutien du Service de la langue française du Ministère de la Communauté française de Belgique.

Programme Committee

T. Fontenelle (European Commission)
M. Gellerstam (Gothenburg)
U. Heid (Stuttgart)
C. Marello (Torino)
A. Michiels (Liège)
A. Moulin (Liège)
S. Theissen (Liège)

Referees' Panel

H. Béjoint, D. Blampain, F. Čermák, M. Devos, A. Duval, T. Fontenelle, D. Geeraerts, M. Gellerstam, G. Grefenstette, U. Heid, T. Herbst, Ph. Hiligsmann, F.E. Knowles, C. Marello, W. Martin, A. Michiels, A. Moulin, R. Moon, O. Norling-Christensen, S. Theissen, K. Varantola, J. Véronis.

We also wish to thank M. Clari, J. Butterfield and A. Reichling, who took charge of the tutorials; Professor Willy Legros, Rector of the University of Liège, who welcomed the participants; A. Cowie, M.H. Corréard and G. Grefenstette who delivered the keynote lectures; all the colleagues who presented papers and posters or agreed to chair our various sessions; and the publishers who took part in the book exhibition. We are also grateful to Nicolas Dufour, who repeatedly lent us a helping hand, and especially to Véronique Doppagne who, day after day, took care of all the practical problems with enthusiastic dedication and unflagging energy.

Thierry Fontenelle, Philippe Hiligsmann, Archibald Michiels, André Moulin, Siegfried Theissen (eds)

ACTES

EURALEX'98

PROCEEDINGS

Communications soumises à EURALEX'98 (Huitième Congrès International de Lexicographie) à Liège, Belgique
Papers submitted to the Eighth EURALEX International Congress on Lexicography in Liège, Belgium

Vol. I

toutefois que si une recherche de haut niveau suppose une certaine spécialisation, elle n'empêche pas pour autant l'interdisciplinarité et des collaborations fructueuses.

Enfin, l'intense activité intellectuelle d'Euralex n'est certainement pas gratuite et ne se limite pas à des débats entre initiés. Au contraire, elle débouche sur la production d'ouvrages de référence mis à la disposition de la communauté tout entière. A cet égard, je tiens à souligner combien notre université est consciente de la dimension citoyenne de ses responsabilités et de la nécessité de rendre le résultat de ses recherches accessible à tous. Outil de précision mais aussi de formation, le dictionnaire doit pouvoir être utilisé tant par l'homme de la rue que par le traducteur technique ou encore par l'étudiant désireux d'améliorer sa connaissance de sa langue maternelle et des langues étrangères. Ceux d'entre nous qui ont charge d'enseignement savent combien sont grands les besoins de nos étudiants dans ce domaine. Un usage adéquat du dictionnaire est précisément à la base de l'acquisition de compétences linguistiques dignes de ce nom. On oublie trop souvent qu'avant de servir sur le plan professionnel, la maîtrise des langues est une condition d'épanouissement, d'ouverture aux autres et d'intégration dans le monde au sens le plus large. Cette réalité donne à ce Congrès une réelle dimension humaniste que j'ai plaisir à souligner.

To conclude, I wish to thank all those who have contributed to the success of this Eighth Euralex Congress and more particularly the Euralex Executive Board, who entrusted our University with its organization.

Our thanks also go to all the organizations which gave us their support and whose names appear in the proceedings, and finally to all the participants who did us the honour of either presenting a paper or attending the sessions.

I wish you all a very fruitful, productive and intellectually rewarding session.

Thank you... and work hard !

Willy LEGROS
Recteur de l'Université de Liège

CONTENTS

The papers of each topic are listed in alphabetical order (except for the plenary lectures that are listed in order of presentation).

VOL. I

Adriana ROVENTINI, Francesca BERTAGNA, Nicoletta CALZOLARI, Istituto di Linguistica Computazionale, CNR
Carol PETERS, Istituto di Elaborazione della Informazione, CNR

Building the Italian Component of EuroWordNet: a Language-specific Perspective

Abstract

The approach being followed to build a semantic database or wordnet for Italian within the framework of the EuroWordNet project is discussed. The emphasis is on the strategies employed to ensure that the monolingual database is linguistically coherent while, at the same time, guaranteeing compatibility with the other components of the project. The paper is divided into two main sections in which we deal with the monolingual and multilingual aspects of the work respectively. In the first part we describe the construction of the core entities of the Italian wordnet - the synsets - and the difficulties encountered when building coherent linguistic/semantic taxonomies. The second part will briefly present the problems faced and the methodology being adopted for a semi-automatic mapping of the Italian lexical data to the Interlingual Index of EuroWordnet.

Keywords: Lexical semantics, EuroWordNet, Italian semantic database, Cross-language mapping

1. Introduction

There is currently much international interest in the potential of WordNet like semantic database systems and a number of initiatives are under way to emulate - to varying degrees - the important work of George Miller and his group in Princeton (Miller et al.: 1990). The aim is to create tools that can be used in different types of language processing tasks, e.g. acquisition of lexical information, sense disambiguation, information retrieval activities. The special feature of EuroWordNet[1], an EC-funded project, is that a set of monolingual semantic nets - in the first phase, Dutch, English, Italian and Spanish - are being linked through an Interlingual Index and thus can also be used for multilingual processing activities such as cross-language information retrieval, contrastive linguistic studies, etc.

In this paper we will not go into details concerning the project as a whole: the interested reader can refer to (Alonge et al.: 1996; Climent et al.: 1996; Vossen: 1997), and to a forthcoming number of *Computers and the Humanities* which will be completely dedicated to EuroWordNet. Our aim here is to describe the particular approach we have taken to ensure that the Italian database is linguistically coherent and that the steps taken to permit cross-language mapping do not obscure or worse eliminate language-specific features.

2. The Italian WordNet

When constructing the Italian WordNet, we had two main concerns: the first was to ensure that the particular features of the Italian lexical system were adequately represented; the second was to guarantee the maximum compatibility with the wordnets being built by the other partners. Our objective was thus twofold: (i) to construct a flexible and useful tool to be

employed in certain Italian NLP tasks, e.g. sense disambiguation; (ii) to create a component for a semantic database that can be exploited in different types of multilingual extraction and analysis activities, e.g. cross-language studies and multilingual information retrieval. In this section, we will discuss the efforts we are making to respect the former commitment, the latter will be discussed in Section 3.

First, however, we must provide some information on two main decisions which characterised the whole project and, consequently, the construction of the Italian wordnet. The first decision was that a vocabulary subset should be selected for each project language. This subset should represent the most general and commonly used word-senses in that language (the criterion being "those 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. The selection of this first set of language-dependent "base concepts" was followed by a stage of cross-language comparison in order to be able to establish a common set of base concepts for all the languages. The second decision was that in EuroWordNet even more attention, with respect to WordNet 1.5, should be given to the notion of the lexicon as a network of relations where any given word-meaning is derived from the set of its relations with other words (see Lyons: 68). Various kinds of semantic relations were thus added to those existing in WN1.5, in particular relations (such as synonymy, antonymy, hyp(er)onymy, meronymy and near-synonymy) between different parts of speech (noun, verb and adjective). In this way, a base concept such as *atto* (act), is related not only to its synonym *azione* (action) and to the set of its hyponyms, but also to the near synonym verb *agire* (to act), and the noun *attività* (activity) is connected with its near synonym adjective *attivo* (active). By means of all these relations, the word-meaning is seen and described from a multiple perspective and can be recognised and identified in many different contextualizations. This decision gives much more strength to the notion of EuroWordnet as a database with a semantically based structure, while facilitating its employment in applications such as information retrieval, in particular in a multilingual environment. A list of the principal internal semantic relations is given in Table 1. For a complete description see (Alonge et al: 1998).

RELATION TYPE	PARTS OF SPEECH
Near Synonym	N<>N, V<>V
XPOS_Near_Synonym	N<>V, N<>AdjAdv, V<>AdjAdv
Has_Hyperonym/Hyponym	N>N, V>V
Has_XPOS_Hyperonym/Hyponym	N>V, N> AdjAdv, V>AdjAdv, V>N, AdjAdv>N, AdjAdv >V
Has_Holonym	N>N
Has_Holo_Part/Member/Portion/Madeof/ Location	N>N
Has_Meronym	N>N
Has_Mero_Part/Member/Madeof/Location	N>N
Antonym/Near Antonym	N<>N, V<>V
Causes	V>V, N>V, N>N, V>N, V>AdjAdv, N>AdjAdv
Is_Caused_by	V>V, N>V, N>N, V>N, AdjAdv>V, AdjAdv>N

RELATION TYPE	PARTS OF SPEECH
Role	N>V, N>N, AdjAdv>N, AdjAdv>V
Role_Agent/Instrument/Patient	N>V, N>N
Role_Location/Direction	N>V, N>N, AdjAdv>N, AdjAdv>V
Involved	V>N, N>N, V>AdjAdv, N>AdjAdv
Involved_Agent/Patient/Instrument	V>N, N>N
Involved_Location/Direction	V>N, N>N, V>AdjAdv, N>AdjAdv
Be_In_State_of	N>AdjAdv, V>AdjAdv
State_of	AdjAdv>N, AdjAdv>V
Eq_Synonym	N<>N, V<>V
Eq_Near_Synonym	N<>N, V<>V, N<>AdjAdv, V<>AdjAdv
Has_Eq_Hyperonym/Hyponym	N>N, N>V, N>AdjAdv, V>V, V>N, V>AdjAdv, AdjAdv>N, AdjAdv>V

Table 1: Major Semantic Relations in EuroWordNet

2.1. Data Sources

An initial decision of the project was to take advantage, as far as possible, of existing tools, methodologies and resources when creating the individual monolingual databases. This decision considerably influenced the approaches taken by the individual partners as data sources and processing strategies differed from site to site.

To preserve language-dependent features, in terms of grouping words through the different relations which reflect language specific interconnections, we decided to start constructing the Italian wordnet from Italian lexical data. The mapping of the Italian net to the English one (and through this to the other languages) was thus performed in a second stage. Furthermore we decided to construct the Italian wordnet from a number of different existing sources available in our Institute (ILC - CNR) in order to be able to overcome, to some extent, the idiosyncrasies of a single dictionary and to provide a more objective perspective on the data. In our opinion this is very important; it is doubtful that a single existing source will be adequate to represent the lexical system of a language. In fact, an integration of different sources has highlighted the differences between dictionaries and the inconsistencies found in dictionary data; e.g. word senses, synonyms, and genus terms can vary widely from source to source. We had four main starting points for our data:

- The Italian Lexical Database (already constructed from a number of sources). The LDB subset used for EuroWordNet currently contains about 30,000 entries (5,500 verbs and 24,500 nouns) totalling about 60,000 word senses. The following semantic relations had already been partially tagged in previous projects (e.g. Acquilex, Delis): synonymy, hyponymy, part-of, set-of, deverbal, deadjectival for nouns; synonymy, hyponymy and causation for verbs.
- An Electronic Synonym Dictionary. This is used as a source for indications on synonym data and word-senses distinctions.
- An Italian/English Lexical Database. This database contains approximately 30,000 senses on each side. It is used to give a first translation of the Italian word-senses and

also as a source of potential synonyms, providing a different perspective from that of the monolingual sources.

- The Italian Reference Corpus, used as an additional source of data, e.g. although multiwords are less common and treated differently in Italian than, for instance, in Germanic languages, we found that they were often important when structuring our semantic hierarchies. However, as they are generally not listed as entries in Italian dictionaries, we needed an objective means to identify them.[2]

The integration of data from these different sources has involved much work. We generally used semi-automatic procedures for a first merging of the data, but a close and careful manual intervention was then necessary to try to make the "right" choice. We will discuss the type of decisions taken in the next three sections with reference to the choice of the core subset and the creation of synsets and lexical semantic hierarchies. The examples given are restricted to nouns, however, with some differences, the procedures followed were much the same for verbs.

2.2. Base Concepts

The core subset of word-meanings was selected from our monolingual lexical database, mainly on the basis of their frequency as genus terms in the definitions, thus ensuring the coverage of most of the other words in the lexicon. A list of about 300 nouns and 100 verbs was extracted and analysed as forming a first set of base concepts for Italian. However, it soon became clear that the number of hyponyms for any genus term could not be a sufficient criterion for the selection of a valid set of base concepts, in part for the simple reason that many important concepts do not have hyponyms. In fact, this preliminary subset was neither homogeneous nor consistent. It reflected strongly the defects and inconsistencies of the lexicographic metalanguage on which it was based and could only be considered as a starting point for the construction of a coherent semantic network. Many concepts were missing and had to be introduced by manual interventions on the data: typical examples are the sets of entries referring to atmospheric phenomena and kinships terms, where we found that simply following the criterion of productive genus terms included terms like wind (in fact there are many different nouns denoting the wind depending on its origin and direction or denoting the typical wind in a specific place or town such as for example *bora* for Trieste or *ponentino* for Roma) but excluded other terms that are intuitively of equal importance, e.g. rain. Further integrations on the data were made on the basis of consultations of other sources (the Italian Reference Corpus for example). A successive step was represented by reference to and subsequent integration of those word-meanings chosen by the other partners which had not been merged during the initial analysis of our data. Table 2 shows the number of synsets proposed by each site and how many of them were selected and how many rejected. The final set of Common Base Concepts consists of 1021 items: 793 Nouns and 228 Verbs.

Nouns	Proposed	Selected	Rejected	Verbs	Proposed	Selected	Rejected
AMS	1027	429	598	AMS	323	126	197
FUE	523	323	200	FUE	128	72	56
PSA	334	239	95	PSA	104	63	41
SHE	1296	594	702	SHE	236	132	104

Table 2: Selected and Rejected Base Concepts over the Project Partner Sites

2.3. Constructing synsets for Italian

In accordance with the WordNet philosophy, where the central semantic relation is that of synonymy, we started the building of our network by searching for the synonyms of the selected base concepts. The project adopted a weak definition of synonymy, entailing the interchangeability of two words in a given context, which could be better denoted as "semantic similarity". This was considered useful to avoid those over granular distinctions which have been observed by many users of WN1.5 as causing problems in applications (e.g. in information retrieval).
For Italian this task was carried out by means of automatic extraction procedures followed by careful manual revisions. The source data for our synsets is a combination of information taken from the sources listed above, using a three-step procedure, as follows:
a) explicitly tagged synonyms in the LDB and in the Synonym Dictionary are grouped to form a first proposal of synset;
b) candidate synonyms, i.e. synonymic type definitions, are associated with all members of the synset under construction;
c) each candidate for the synset is searched in the Bilingual Dictionary: semantic indicators and translation equivalents are associated.

When revising these automatically created synsets we found that a sense shifting often occurs. This phenomenon is unavoidable and must be controlled. Very often the synsets appear too large and manual revision is necessary to cut the automatically associated synonyms groups according to more coherent boundaries. To give just an idea, we show the results of this procedure when building a synset for the concept represented by the Italian word *ansia* (anxiety). The automatic extraction of synonyms gives us a very large set of candidates for this synset:

> *ansia, ansietà, affanno, ambascia, travaglio, timore, inquietutine, pena, apprensione, trepidazione, angoscia, dolore, tormento, afflizione, strazio, patimento, tristezza, accoramento, supplizio, sofferenza, malinconia, martirio, tortura.*

However, this example shows why revision of even explicitly tagged dictionary synonyms is necessary. At a certain point, within the synset, a twofold meaning shift occurs moving from the general idea of anxiety to either that of anguish / suffering (represented by *angoscia, dolore, tormento, strazio...*) or melancholy / sadness (*tristezza, accoramento, malinconia*). The synonym chain is thus interrupted manually. The final synset for ansia (anxiety) was: {*ansia, ansietà, inquietudine, pena, preoccupazione, apprensione*}.
When grouping our synsets we must keep our final goal firmly in mind: to build a truly representative lexical/semantic network while providing a useful tool for language processing and information retrieval activities. Ideally our synset should be sufficiently extensive to embrace a concept lexically (high recall) but not so loose as to include scarcely related concepts (low precision).

2.4. Constructing the Semantic Hierarchies

Once our base concepts were structured in synsets, integrated, and linked to WN1.5 by means of a careful manual operation (see Section 3 below), the top-down extension of the taxonomies was carried out using a semi-automatic procedure to retrieve all hyponyms for each synset or word-meaning. Starting from our automatically created noun taxonomies, a

difficulty we had to face was again the inconsistencies derived from the definitions themselves. To give a concrete example, we can examine the taxonomy of *strumento* (instrument, tool) which is one of the most important base concepts. In this taxonomy, which contains about 1,000 lexical items, we found phenomena such as: (i) circularity in the definitions of the top concepts, which means that we had to find a suitable criterion to decide the right hyperonym / hyponym relations within the taxonomy; (ii) many different types of hyponyms for the same hyperonym; (iii) different genus terms used for identical types of objects.

In Italian the most general and comprehensive word for the English "instrument/tool" is *strumento* and this is actually the most frequent genus term in this field, being used to define 290 items. Unfortunately, we must address two basic problems that are caused by the inconsistency of the definitions in our main source (the LDB): (a) *strumento* has been assigned as hyperonym the word *arnese* which is not perceived as more general; (b) *strumento* has only two word senses, the first covering all its concrete meanings, the second the figurative and extended ones.
The first point gives rise to a problem of circularity because *arnese* has as synonym *utensile* and hyperonyms *attrezzo* or *strumento*; while, in its turn, *attrezzo* has as hyperonyms *arnese* and *strumento*, and finally *utensile* has as hyperonym *arnese*. This circularity determines (and can be considered as a proof of) a first synset: {*strumento, arnese, attrezzo, utensile*}. But, if we consider the more general use of *strumento* and also its possibility of being employed in figurative and extended senses we should place this word on a higher level (compared with the other three) within the taxonomy. In fact, in Italian, we can define as *strumento* nearly all types of tools, but the same is not true for *arnese* or *attrezzo* or *utensile* which have a narrower denotation:

la zappa é uno strumento (the hoe is a *strumento*)
il computer é uno strumento (the computer is a *strumento*)
la zappa é un arnese / un attrezzo (the hoe is an *arnese / attrezzo*)
* *il computer é un arnese / un attrezzo* (the computer is an *arnese / attrezzo*)

We also find that very different types of instruments were listed under this genus: we found simple manual instruments, scientific measuring instruments and musical instruments mixed together, i.e. here we have a typical example of under-differentiation between word senses. In this and in similar cases we need a finer-grained distinction with respect to our sources, giving rise to a greater number of sub-taxonomies, based on other features which are found in the "differentia" part of the definitions.
The last problem to be observed with this particular (but typical) taxonomy was concerned with the different genus terms used to define strongly related objects such as, for example, pieces of cutlery. Examining the data we found *forchetta* (fork) under *arnese*, but *coltello* (knife) and *cucchiaio* (spoon) are found under *strumento* and *posata* respectively. For cases like this, we need to correct the incoherence by using the appropriate level in the taxonomy for all the related words, i.e. the lowest appropriate level (in this case *posata*, which in turn will point to *utensile* and thus to *strumento*).

As can be seen, the work of restructuring the taxonomies required much manual intervention to add intermediate levels for large sets of hyponyms, where many very specific terms were directly linked to generic hyperonyms at a too high level. In the instruments taxonomy, we introduced multiwords, which do not appear as lexical entries in the Italian monolingual LDB, but are lexicalized expressions (in keeping with the decision of building a lexical net in

EuroWordNet rather than a conceptual net) such as *strumenti musicali* (musical instruments), *strumenti di misura* (measure instruments). In this way, we created a new level in the taxonomy and, at the same time, more homogeneous lexical subsets. Another typical example of this is constituted by the "person" taxonomy where concepts such as *artista* (artist), *lavoratore* (worker), *seguace* (follower) etc., have been introduced as an intermediate level between generic and specific concepts. So now we have:

{*persona, essere umano, individuo, uomo*} (person, human being, individual, man) as our base synset,
artista, lavoratore, seguace, ... (artist, worker, follower,..) first level hyponyms
musicista, pittore, scrittore,.. (musician, painter, writer,..) second level (hyponyms of *artista*)
pianista, sassofonista,.. (pianist, saxophonist,..) third level (hyponyms of *musicista*)

whereas previously the taxonomy went directly in one step from pianist to person.

3. Mapping to the Interlingual Index

In EuroWordNet, all the language specific wordnets will be stored in a central lexical database system. Equivalence relations between the synsets in different languages will be made explicit through an Interlingual Index (ILI). This will be a modified but unstructured version of WN1.5 in which original senses will be modified and new senses added if necessary. Each synset in the monolingual wordnets will have at least one equivalence relation with an ILI record which will enable cross-language mapping and comparison. This can be an equivalent synonym relation when there is an exact matching between the Italian and English data (e.g. *animale* matches exactly to animal), an equivalent near-synonym relation when the match is close but not precise (e.g. *polpetta* is matched as equivalent near synonym to rissole, the concept is the same but the realisation is different) and an equivalent hyperonym relation when we are dealing with language specific objects that have no match in the other language (e.g. the Italian cake made from chestnut flour *castagnaccio* is linked to cake with an equivalent hyperonym relation). Linked to the ILI is a language independent Top Ontology and a set of domain labels.

We have developed a semi-automatic procedure to establish these equivalence relations between the Italian data and WordNet synsets. This is not simple. 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 will be unstructured, we have exploited the structure of WN1.5 in order to make the right connections between the Italian lexical entries and the WN senses.

Our mapping procedure operates taxonomy by taxonomy. We start with the base concepts that had already been mapped manually to our ILI through WN1.5 and therefore provide us with a set of accurate anchor points between the Italian database and WN1.5. 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 a word-form that matches one of the candidate translations; the assumption is that matching word-forms in equivalent semantic hierarchies in different languages will refer to equivalent senses.

The results of the automatic stage of the mapping procedure then have to be checked and integrated manually in a second stage. At the end of the first stage we have four possible results: (i) unambiguous mapping to an equivalent WN1.5 sense; (ii) more than one possible mapping proposed; (iii) a bilingual translation but no WN1.5 equivalent; (iv) no bilingual translation found and thus mapping with a has-equivalent_hyperonym relation to the WN1.5 equivalent base concept. In the manual revision stage, we have to evaluate and resolve cases ii, iii, and iv. Frequently has-equivalent_near_synonym and intermediate has-equivalent_ hyperonym relations are introduced when no exact equivalent can be found.

The main problems we encountered in matching to WordNet were differences in lexicalization, mismatches and lexical gaps. We give here a few examples of these difficulties (and consequent issues raised and solutions devised):

(i) Very frequently the Wordnet distinctions are too fine-grained - it appears that the Italian item could match equally well to more than one level of a given taxonomy, e.g. for *stabilmento* which was translated by our bilingual LDB as plant, factory, it is not easy to decide whether it is best linked to WN1.5 {factory, mill, manufacturing plant, manufacturing} or to its direct hyperonym {plant, works}. Cases like this suggest a possible merging of the relevant senses of WN1.5.

(ii) Similarly, it often occurs that a single Italian item can match equally well to more than one WordNet synset. For example, we have *oggetto* 2 which maps to both {aim, object, objective, target} and also {purpose, intent, intention, aim}. As these two synsets both belong to the same taxonomy (which terminates in {psychological feature} passing via {goal, end}), it appears reasonable again to propose a merging between the WN1.5 items for our Interlingua. However, a proposal of this type is probably not feasible when our Italian item matches to WN1.5 entries which belong to different taxonomies, e.g. we have *stato* 4, translated by the bilingual LDB into state, which has currently been mapped as equivalent near synonym to three WN1.5 entries: {state, province, territory}, {country, state, land, nation}, {state, nation, country, land, commonwealth, res pubblica, body politic}. In this case, the first WN1.5 entry is in the location taxonomy, whereas the other two belong to {group, grouping}. This suggests that probably the Italian entry should be revised and perhaps split into two senses.

(iii) Indeed, frequently a single sense in the Italian data is already clearly split by our bilingual LDB into more than one sense; in such cases again we create separate senses in the Italian WordNet. An example of this is *macchina* 1 which in fact encapsulates the very different senses of machine, engine and car; the cross-language mapping thus suggests that we should reconsider our original encoding of *macchina* in the Italian wordnet and split it into three separate synsets, eg. {*macchina, motore*}, {*macchina, locomotiva*} and {*macchina, automobile*}.

(iv) Another, less frequent case, is when we can find no Wordnet equivalent sense, e.g. the Italian *elenco* is naturally translated as list, in the sense of number of items written or printed; however, the only WN1.5 entry under list is list, listing (glossed as a database ...). This is clearly a limited sense of list and not that implied by *elenco*. In these cases, presumably, we must add a new sense to our Interlingua.

(v) Finally when it is not possible to establish a direct equivalent near synonym relation between our data and an ILI record, we use the equivalent near synonym or equivalent hyperonym relations. 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-equivalence_hyperonym relation. On the contrary, relations of equivalent hyponymy will be established between Italian *dito* and the ILI records for finger and toe.

As can be seen the cross-language mapping stage also provides useful insight and feedback on the structuring and coherency of the monolingual database. It gives us the opportunity to verify the Italian data and, when necessary, to restructure it or complete it when lexical gaps are evidenced.

4. Final Remarks

Much attention is being currently paid by international research community to the potential of Wordnet-like semantic databases for many types of applications (e.g. mono- and multilingual IR activities). This has led to the consequent interest in the construction of such resources. In the design phase of EuroWordNet we have taken a number of decisions aimed at enabling the use of the resource in many applications (also on the basis of experiences by other groups in using the existing WN1.5). Among these choices we mention: less fine-grained sense distinction, a common shared Top-ontology, a comparable (cross-linguistically) set of base concepts, a larger set of relations (also between different POSs), an Interlingual Index (to map between the various languages). These strategic decisions have obvious consequences on the methodology of work, and raise challenging problems while building the resource both at monolingual and at multilingual levels. Our objective has been to construct a linguistically coherent semantic net for Italian which can be used in Italian NLP tasks while, at the same time, is compatible and consistent with the overall design of the multilingual database.

5. Notes

1 The project (LE4003) partners are currently: University of Amsterdam (coordinator), Fundacion Universidad Empresa (a cooperation of UNED Madrid, Politecnica de Catalunya, Barcelona, and University of Barcelona), University of Sheffield, Istituto di Linguistica Computazionale, CNR, Pisa and Novell Linguistic Development (Antwerp). In a second stage, the database should be extended with German, French, Estonian and Czech.

2 This is important. The EWN databases are lexical rather than conceptual nets; this means that each entry must be recognized as a lexical item in that language.

6. References

Alonge, A. (1996). "Definition of the links and subsets for verbs", EuroWordNet Project LE4003, Deliverable D006. Http://www.let.uva.nl/~ewn.

Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, T., Peters, W. (forthcoming). "The Linguistic Design of the EuroWordNet Database", to appear in *Computers and Humanities*, 1998.

Climent, S., Rodríguez H., Gonzalo J. (1996). "Definition of the links and subsets for nouns of the EuroWordNet project", EuroWordNet Project LE4003, Deliverable D005. Http://www.let.uva.nl/~ewn.

Lyons, J. (1968). "*Introduction to Theoretical Linguistics*. Cambridge University Press.

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.

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

VOL. II