LANGUAGE INTERNAL RELATIONS FOR NOUNS, VERBS, ADJECTIVES AND ADVERBS

Deliverable D.5.1, WP5, BalkaNet,
IST-2000-29388

BalkaNet

Identification number	IST-2000-29388
Type	Report – Document
Title	Language Internal Relations for Nouns, Verbs, Adjectives and Adverbs
Status	Final
Deliverable	D.5.1
WP contributing to the deliverable	WP5
Task	T.5.1
Period covered	July 2002 – November 2002
Date	12 December 2002
Status	Confidential
Version	1
Number of pages	38
WP/Task responsible	SABANCI
Other Contributors	DBLAB, UAIC, RACAI, CTI, FI MU, MATF, DBMB, PU, UOA
Authors	{Dan Cristea, Georgiana Puscasu, Oana Postolache} **UAIC** {Sofia Stamou} **DBLAB** {Cvetana Krstev, Gordana Pavlovic-Lazetic, Ivan Obradovic, Dusko Vitas} **MATF** {Orhan Bilgin, Ozlem Cetinoglu} **SABANCI** {Dan Tufis} **RACAI** {Karel Pala, Pavel Smrz} **FI MU** {Svetla Koeva} **DCMB**

Project Coordinator	Professor Dimitris Christodoulakis Director of DBLAB Databases Laboratory, Computer Engineering & Informatics Department Patras University GR 26500, Greece Phone: +30 (61) 960 385 Fax: +30 (61) 960 438 E-mail: dxri@cti.gr
EC project officer	Erwin Valentini
Actual distribution	Project Consortium, Project Officer, EC
Key words	Semantic relations, language-specific relations, monolingual wordnets, morphology
Abstract	This deliverable concerns the semantic relations of each monolingual wordnet of the BalkaNet Project (Bulgarian, Czech, Greek, Romanian, Serbian, Turkish). As the project aims at integrating the final wordnets not only with each other but also with EuroWordNet, the basic relations of synonymy and hyperonymy/hyponymy have been implemented in each wordnet. The other relations specified in EuroWordNet will also be used where possible. On the other hand, new language-specific relations should be defined since the morpho-semantic structures of the languages differ from each other. Each partner has provided a detailed description of such relations in this Deliverable. These new relations are proposals at the moment and should be revised by all the members of the consortium before implementation.
Status of the abstract	Complete
Send on	12 December 2002

EXECUTIVE SUMMARY

This document reports on the semantic relations within the BalkaNet Project which have been encoded so far and will be encoded at the later stages of the project.

Except for the Czech team, each partner is developing its wordnet from scratch. To have a common frame, the BalkaNet consortium agreed on translating the BCs from English into the participants' languages, which will ensure that the mainframe of each wordnet is constructed in the same manner. As a next stage, the participants proposed different sets of concepts and the consortium agreed upon a set called Subset 2 to continue with.

The wordnets developed by now do not only contain translated versions of the Base Concepts and Subset 2 but have also implemented the semantic relations which have been inherited from EuroWordNet.

The basic relations synonymy and hyperonymy/ hyponymy have been implemented by all the partners. The remaining relations implemented are given below, together with the name of the participant which has implemented it (the figures have been taken from the respective sections prepared by each partner).

- Bulgarian

Be_In_State, Causes, Holo_Member, Holo_Part, Holo_Portion, Near_Antonym, Near_Synonym, Subevent

- Romanian

Be_In_State, Near_Antonym, Subevent, Eq_Hyperonym, Near_Synonymym, Hyperonym, Causes, Holo_Portion, Holo_Part, Eq_Hyponym

- Serbian

Holo_Part, Eq_Generalization, Holo_Member, Near_Antonym, Subevent, Causes, Near_Synonym, Eq_Metonym, Eq_Diathesis, Be_In_State, Holo_Portion

- Turkish

Holo_Part, Eq_Generalization, Holo_Member, Near_Antonym, Subevent, Causes, Near_Synonym, Eq_Metonym, Eq_Diathesis, Be_In_State, Holo_Portion

The partners have also defined relations which are necessary for representing their language-specific morphology and morpho-semantics.

The Bulgarian team explains the multiple hyperonymy relation with instances taken from English WordNet. Multiple hyperonymy relations should be implemented not only by the Bulgarian team but also by all other members of the Consortium. Indeed, this relation has already been implemented in EuroWordNet for a small number of instances but has not been generalized.

For the Czech partners, relations for verb aspect (imperfectives, perfectives, iteratives), reflexive verbs, verb prefixation (single, double), diminutives (noun derivation by

suffixation), move in gender (noun derivation by suffixation), and other types of word form derivation (word derivation nests, families) should be defined and implemented besides already defined eq_synonym and near_synonym relations.

The Greek team proposes to implement the so-called "belong_to" relation to link each concept to its domain, which will be rather useful for information retrieval and conceptual indexing purposes. Additionally, the antonymy, meronymy/holonymy, role/agent, involved agent, role/instrument, involved patient, and derived_from relations, which are defined in EuroWordNet, should also be implemented

The Serbian team mentions relations involving a so-called structural derivation, where the meaning of a derived word can be predicted from the original word and the derivational process. The Serbian explanatory dictionary is suitable for extracting such relations automatically.

The Turkish partners define several morpho-semantic relations and also propose a structural change in the existing entry format, in order to represent purely morphological relations between word forms. Since Turkish is an agglutinative language, there are several highly productive derivational suffixes, some of which lead up to automatically encoded relations. Some of the relations are not so productive and have a limited number of instances and should therefore be added manually, if they are to be implemented. It is also possible to define a relation denoting etymological information of concepts.

TABLE OF CONTENTS

1. INTRODUCTION 8
2. RELATIONS ALREADY IMPLEMENTED 9
2.1. BULGARIAN WORDNET 9
2.1.1. SYNONYMY 10
2.1.2. HYPERONYMY / HYPONYM 11
2.1.3. BE_IN_STATE / STATE_OF 11
2.1.4. CAUSES / IS_CAUSED_BY 12
2.1.5. MERONYM 12
HOLO_MEMBER / MERO_MEMBER 12
HOLO_PART / MERO_PART 12
HOLO_PORTION / MERO_PORTION 13
2.1.6. NEAR_SYNONYM 13
2.1.7. NEAR_ANTONYM 13
2.1.8. HAS_SUBEVENT / IS_SUBEVENT_OF 14
2.2. CZECH WORDNET 14
2.2.1. Verb Valencies and Verb Senses 14
2.3. GREEK WORDNET 15
2.3.1. Common Language Internal Relations 15
2.3.2. Language Internal Relations Encoded So Far in Greek WordNet 16
SYNONYMY 16
HYPONYMY / HYPERNYMY 17
2.4. ROMANIAN WORDNET 17
2.5. SERBIAN WORDNET 19
2.6. TURKISH WORDNET 19
2.6.1. CAUSES and IS_CAUSED_BY Relations in Turkish 20
2.6.2. BE_IN_STATE and STATE_OF Relations in Turkish 21
3. DISCUSSION ON THE SEMANTIC RELATIONS THAT CAN BE ENCODED 22
3.1. BULGARIAN 22
3.1.1. Multiple Hyperonymy Relations 22
3.2. CZECH 23
3.2.1. Aspect Opposition 24
3.2.2. Reflexives 24
3.2.3. Prefixed verbs 25
3.2.4. Gender pairs 25
3.2.5. Diminutives 25
3.3. GREEK 26
3.3.1. Belongs to 26
3.3.2. Antonymy 26
3.3.3. Meronymy / Holonymy 26
3.3.4. Derived_from 27
3.3.5. Role/Agent, Involved Agent, Role/Instrument, Involved Patient 27
3.4. ROMANIAN 27
3.5. SERBIAN 27
3.6. TURKISH 28
3.6.1. Word-to-Word Relations vs. Synset-to-Synset Relations 28
3.6.1. INVOLVED_X Relations and CO_X_Y Relations 30

3.6.2. Verbs Derived From Nouns and Adjectives Using The Support Verbs “Etmek” and “Olmak” 32
3.6.3. FUZZYNYM and XPOS_FUZZYNYM 32
3.6.4. “Become Verbs” Derived From Nouns and Adjectives Using the Suffix “-laşmak” 32
3.6.5. Verbs Derived from Nouns and Adjectives Using the Suffix “-lamak” 32
3.6.6. Verbs Derived From Nouns Using the Suffix “-lanmak” 33
3.6.7. Reflexive Verbs Derived From Action Verbs Using the Suffix “-(A)Nmak” 33
3.6.8. Classes and Instances 33
4. TECHNIQUES USED FOR EXTRACTING RELATIONS 33
4.1. BULGARIAN 33
4.1.1. Automatic Translation 33
4.1.2. Automatic Assignment of Additional Synonyms 34
4.1.3. Automatic Assignment of Hyperonyms 34
4.2. GREEK 34
4.2.1. Extraction of Language Internal Relations for Greek 34
4.2.2. Problems Encountered during Selection Process of the Greek Base Concepts 35
4.3. SERBIAN 36
4.4. TURKISH 36
4.4.1. Synset Extraction from TDK (Turkish Monolingual Dictionary) 36
4.4.2. Extraction of Other Relations from TDK 37
HYPERONYMY 38
ANTONYMY 38
5. CONCLUSION AND FUTURE WORK 39

1. INTRODUCTION

The work reported in this deliverable is the output of the first task (T.5.1) of Work package (WP) 5 which defines existing semantic relations or those which will be implemented by the end of the project, of the monolingual WordNets for each of the languages participating in the BalkaNet project (Bulgarian, Czech, Greek, Romanian, Serbian and Turkish).
The language internal relations of the BalkaNet project will be mostly based on the relations adopted in the Princeton WordNet 1.5 and the EuroWordNet projects. The motivation for deciding to adopt already existing relations is twofold. First and foremost we wish to keep compatibility with existing WordNets in terms of structure of the lexical items and secondly due to the fact that most of the existing relations are capable of declaring in an efficient way semantic relationships between terms and concepts.

Moreover, from an Information Retrieval (IR) point of view and with respect to the envisaged application of BalkaNet's results some of the already existing relations (e.g. synonymy, hyponymy and hyperonymy) are rather useful for retrieval of information. Consequently, as already been decided within the consortium in each monolingual WordNet concepts will be inter-linked on the basis of the language internal relations adopted both in EWN and Princeton WordNet.

However, some differentiations might come up due to language particularities and the quality and structure of the available lexical resources for the Balkan languages. In particular, during processing Balkan lexical resources the need for the inclusion of additional language internal links might come up due to the fact that the languages we are dealing with are extremely rich in morphology and thus have a high degree of morphological complexity.

In cases where the need for the introduction of a new relationship comes up by one or more partners such a possibility will be discussed and common, coherent decisions will be made.

The report first describes the relations that have already been implemented in each individual wordnet, in Section 2, and then describes the new relations proposed by each participant, in a separate section (Section 3). Section 4 describes the techniques used for extracting the relations. The last section, Section 5, summarizes the conclusions we made and the future work we plan to perform.

2. RELATIONS ALREADY IMPLEMENTED

Relations implemented in individual wordnets of the BalkaNet Project differ from each other due to differences between the approaches used in constructing the relations of the monolingual wordnets. But they have a common set of relation types which are taken over from EuroWordNet. The basic relations synonymy and hyperonymy/hyponymy have been implemented by all partners. Other relations that have been implemented vary both in terms of their type and the number of occurrences.

2.1. BULGARIAN WORDNET

We started with the encoding of synonymy and hyperonymy / hyponymy language internal relations. At the moment Bulgarian WordNet consists of 5119 synsets: 1310 Base Concepts (Subset 1) of the EuroWordNet Project, 3,690 Base Concepts selected by the partners in the BalkaNet Project (Subset 2) and additional 119 synsets.

The relations in Table 2.1 (other than synonymy and hyperonymy / hyponymy) are included in the corresponding English 5119 synsets from the WordNet 1.5.

RELATION TYPE IN EW1.5	NUMBER OF OCCURRENCES
BE_IN_STATE	180
CAUSES	63
EQ_DIATHESIS	5
EQ_GENERALIZATION	2726
EQ_METONYM	86
HOLO_MEMBER	134
HOLO_PART	333
HOLO_PORTION	22
HYPERONYM	4899
ILI	5119
LITERAL	12533
NEAR_ANTONYM	733
NEAR_SYNONYM	144
POS	5119
SENSE	12533
SUBEVENT	293
SYNONYM	5119
SYNSET	5119

Table 2.1: Relation types and number of occurrences in EWN 1.5

We assigned all relations included in English WordNet 1.5 synsets (equivalent with 5119 Bulgarian entries) to Bulgarian WordNet. Afterwards we excluded those relations that refer to the synsets not yet implemented in Bulgarian WordNet. Then a linguist verified the validity of additional relations, such as HOLO_MEMBER, HOLO_PART, HOLO_PORTION, NAER_ANTONYM, NEAR_SYNONYM, SUBEVENT, BE_IN_STATE and CAUSES. After correspondence checking and verification we achieved the results given in Table 2.2 for Bulgarian.

RELATION TYPE IN BW	NUMBER OF OCCURRENCES
BASE	5119
BE_IN_STATE	10
CAUSES	42
GLOSS	5119
HOLO_MEMBER	38
HOLO_PART	156
HOLO_PORTION	7
HYPERONYM	4898
ILI	5119
LITERAL	10759
NEAR_ANTONYM	237
NEAR_SYNONYM	5
POS	5119
SENSE	10759
SUBEVENT	49
SYNONYM	5119
SYNSET	5119

Table 2.2: Relation types and number of occurrences in BWN

2.1.1. SYNONYMY

Synonymy is symmetric, transitive relation of equivalence. In the first stage of our work we were trying to apply the true synonymy only. This relation implies that the synonyms may substitute vice versa in every context. That is why in Bulgarian data base the average number of synonyms in a synset is 2.1.

Example from Bulgarian WN:

a) If it is a ‘църква’ (church) then it is a ‘черква’ (church)
b) If it is a ‘черква’ (church) then it is a ‘църква’ (church)

<SYNONYM><LITERAL>черква</LITERAL><LITERAL>църква</LITERAL></SYNONYM>

Bulgarian WordNet consists of 10 759 different LITERALS organized in 5 119 SYNSETS.

It is important fact that most ‘synonyms’ have some specific properties, which do not make them fully substitutable. It is a good idea to include in synsets operators marking:

- differences in style and register;
- differences in non-functional grammatical properties (syntactic gender, plurale tantum);
- differences in functional grammatical properties (valency);
- differences in collocations; etc.

2.1.2. HYPERONYMY / HYPONYM

Hyperonymy and hyponymy are inverse, asymmetric and transitive relations, which correspond to the notion of class-inclusion:

If W1 is a kind of W2, then W2 is hyperonym of W1 and W1 is a hyponym of W2.

The relation implies that the hyperonym may substitute the hyponym in a context but not the other way about.

Example from Bulgarian WN:

a) А ‘ястие’ (dish) is a kind of ‘ядене’ (food)
b) ?А ‘ядене’ (food) is a kind of ‘ястие’ (dish)

‘ястие’ (dish)	HAS_HYPERONYM	‘ядене’ (food)
‘ядене’ (food)	HAS_HYPONYM	‘ястие’ (dish)

HYPERONYMY relation is encoded in 4.898 synsets in Bulgarian WordNet.

2.1.3. BE_IN_STATE / STATE_OF

This relation encodes links between nouns that refer to anything in a particular state expressed by an adjective.

Example from Bulgarian WN:

a) ‘големина’ (size) is something to which the state ‘голям’ (big) applies

‘големина’ (size) N	BE_IN_STATE	‘голям’ (big, large) A
‘голям’ (big, large) A	STATE_OF	‘големина’ (size) N

BE_IN_STATE relation is encoded in 10 synsets in Bulgarian WordNet– the parallel synsets from the English WordNet for which we have verified the correspondence in Bulgarian.

2.1.4. CAUSES / IS_CAUSED_BY

The causal relation is used to link verbs, nouns and adjectives with the constraint that the causing event should be dynamic, whereas the resulting situation can either be static or dynamic.

Example from Bulgarian WN:

a) ‘свиквам (среща)’, ‘събирам’ (call) causes ‘събирам се, ‘срещам се’ (gather, meet)

*b) ‘събирам се, ‘срещам се’ (gather, meet) causes to ‘свиквам (среща)’, ‘събирам’ (call)

‘свиквам (среща)’, ‘събирам’ (call) CAUSES ‘събирам се, ‘срещам се’ (gather, meet)
‘събирам се, ‘срещам се’ (gather, meet) IS_CAUSED_BY ‘свиквам (среща)’, ‘събирам’ (call)

We have 42 cases of CAUSES relation encoded in Bulgarian WordNet. The relation could be further specified as non-factive CAUSE.

2.1.5. MERONYM

Part-Whole relation is a family of relations. There are three different types of meronymy relation illustrated in Bulgarian WordNet.

HOLO_MEMBER / MERO_MEMBER

This is a relation between a set and their members.

Example from Bulgarian WN:
a) ‘човешко същество’ (human) is a member of ‘хора’ (people)
***b)** ‘хора’ (people) is a member of ‘човешко същество’ (human)
‘човешко същество’ (human) HAS_HOLO_MEMBER ‘хора’ (people)
‘хора’ (people) HAS_MERO_MEMBER ‘човешко същество’ (human)

There are 38 cases of HOLO_MEMBER relation encoded in Bulgarian WordNet.

HOLO_PART / MERO_PART

This is a relation between (the nouns standing for) a whole and their constituent parts.
Example from Bulgarian WN:

a) ‘покрив’ (roof) is a component of ‘сграда’, ‘здание’ (building)
***b)** ‘сграда’, ‘здание’ (building) is a component of ‘покрив’ (roof)

‘покрив’ (roof) HAS_HOLO_PART ‘сграда’, ‘здание’ (building)
сграда’ (building) HAS_MERO_PART ‘покрив’ (roof)

There are 156 cases of HOLO_PART relation encoded in Bulgarian WordNet.

HOLO_PORTION / MERO_PORTION

This is a relation between a portion and the whole from which it has been detached.

Example from Bulgarian WN:

a) 'телесна или растителна тъкан' (tissue) is a component of 'живо същество, организъм' (life form, organism)
***b)** 'живо същество, организъм' (life form, organism) is a component of 'телесна или растителна тъкан' (tissue)

'живо същество, организъм' (life form, organism) HAS_HOLO_PORTION 'телесна или растителна тъкан' (tissue)
'телесна или растителна тъкан' (tissue) HAS_MERO_PORTION 'живо същество, организъм' (life form, organism)

There are 7 cases of HOLO_PORTION relation encoded in Bulgarian WordNet.

2.1.6. NEAR_SYNONYM

In many cases there is a close relation between words but not sufficient to make them members of the same. For these cases the NEAR_SYNONYM relation is used.

Example from Bulgarian WN:

'странен, необичаен' (bizzare, eccentric) NEAR_SYNONYM 'необикновен, оригинален' (unconventional)
'необикновен, оригинален' (unconventional) NEAR_SYNONYM 'странен, необичаен' (bizzare, eccentric)

NEAR_SYNONYM relation is encoded in 5 synsets in Bulgarian WordNet.

2.1.7. NEAR_ANTONYM

Antonymy is considered to be a relation between word forms, but not between word meanings. If the antonymy relation holds between all variants, the relation is NEAR_ANTONYM, otherwise it is ANTONYMY.

Example from Bulgarian WN:
'продавам' (sell) NEAR_ANTONYM купувам' (buy),'закупувам' (buy)

There are 237 cases of NEAR_ANTONYM relation encoded in Bulgarian WordNet.

2.1.8. HAS_SUBEVENT / IS_SUBEVENT_OF

The relation SUBEVENT is applied to those cases that cannot be expressed by the more specific hyponymy and cause relations.

Example from Bulgarian WN:
a) 'търгувам' (deal in, merchandise, trade) takes place during or as part of 'купувам, акупувам' (buy, purchase)
***b)** 'купувам, акупувам' (buy, purchase) takes place during or as part of 'търгувам' (deal in, merchandise, trade)

'търгувам' (deal in, merchandise, trade) IS_SUBEVENT_OF 'купувам, акупувам' (buy, purchase)
'купувам, акупувам' (buy, purchase) HAS_SUBEVENT 'търгувам' (deal in, merchandise, trade)

There are 49 cases of SUBEVENT relation encoded in Bulgarian WordNet.

2.2. CZECH WORDNET

2.2.1. Verb Valencies and Verb Senses

When building Czech verb synsets we are paying a systematic attention to the verb valencies in a more detailed way than it is done in EWN 1, 2. This follows from inflectional nature of Czech which displays a rich declension structure – each Czech noun (as well as adjective, pronoun and numeral) can appear in one of seven surface cases: `Nominative, Genitive, Dative, Accusative, Vocative, Locative` and `Instrumental`.

For verbs it means that their arguments (participants) represented by nouns or noun groups also come in the mentioned cases: though we can take advantage of the Roles as they are introduced in EuroWordNet 1, 2 for Czech verbs we have to fill in the links between the Roles (deep cases) and seven morphological (surface) cases existing in Czech. This is done by means of valency frames, i.e.each Czech verb synset (or more precisely, each literal in a synset) is linked to its respective valency frame displaying the information about the corresponding deep and morphological cases that are obligatorily (or optionally) associated with it.

In fact, it should be noticed that WordNet 1.5 contains only the surface valency frames (not the Roles), and EWN database on the contrary comprises only the Roles as defined in ILR table above but not the links to the surface cases. In our view for the inflectionally rich languages included in BalkaNet this issue will have to be solved in a more integrated way.
The solution that is applied in Czech WordNet as it is developed now takes the following form (examples are given for verb *držet* and some of its senses):

[v] *držet:5*
ILI: 01500117-v [v] keep:15.
Gloss: "He is keeping three women in the guest cottage"
Valency: kdo*ACT =koho*PAT <kde>*LOC zloděje ve vězení

[v] *držet:11, světit:1.*
ILI: 01458359-v [v] celebrate:3, keep:13, observe:8
Gloss: of holidays or rites; "Keep the commandments"
Valency: {držet, světit} kdo*ACT =co*PAT svátky

[v] *svírat:1, držet:1, udržet:1, podržet:1, třímat:2.*
ILI: 00692795-v [v] hold:13, take hold:2.
Gloss: have or hold, e .g. in one 's hands; "Hold this bowl for a moment, please"
Valency: {mít, třímat} kdo*ACT= co*OBJ v čem*MEANS d. knihu v ruce
Valency: {svírat, držet} kdo*ACT= co*OBJ (v čem*MEANS|[čím]*MEANS) d. volant rukama

The presented notation captures:

- links between surface and deep cases: kdo*ACT, where 'kdo' refers to the surface case Nominative and *ACT represents the fact that the Nominative is associated with the deep case (role) ACTOR (or Agent),
- typical complements that occur with a given verb – they take the form of examples,
- links between the respective sense and the corresponding valency – it can be seen that the different senses are associated with selected valency frames and not with all of the possible ones that may linked to a given literal.

In our opinion this solution may be interesting also for other languages within BalkaNet that display a rich morphological structures (cases or their analogs).

2.3. GREEK WORDNET

2.3.1. Common Language Internal Relations

The relations and the coverage to be represented in the BalkaNet semantic network are described having in mind what is required for the specific end users of the project's results and for the application of the BalkaNet database, given the state of the art in semantics (Princeton WordNet 1.5 and EuroWordNet), the quality of the lexical resources and what is feasible given the available resources, tools and time. Towards this direction the consortium has decided that a minimum set of relations has to be common and reflected in all monolingual lexical networks and the Inter-Lingual-Index. These relations are hypernymy, hyponymy and synonymy. Apart from having a coherent and common set of language internal relations among all WordNets another reason for deciding to represent the aforementioned relations in the final database was the fact that we wanted to achieve a minimum degree of compatibility with the EuroWordNet lexical database since once the project is finished an attempt will be made in order to unify both semantic networks in one common European WordNet. Finally, the reason for concluding on the abovementioned types of relations is that they are common across all languages and they can be extracted from the lexical resources that are already available to the consortium. Furthermore there is coherence among the linguistic community of what these relations stand for and what type of link they denote between two concepts.

However, since we wish to develop WordNets that are representatives of the underlying languages and that include as many terms of the generic vocabulary as possible it is easily understood that in the monolingual WordNet the abovementioned relation types might not be always sufficient to denote the link among terms and that new relations will have to be introduced, which might be common with other languages or not. In light of the above the consortium decided that the three language internal relations i.e. synonymy, hypernymy and hyponymy are going to be present in all monolingual WordNets and in the Inter-Lingual-Index whereas new relation types will be included in the monolingual WordNets. In principle only these relations will be expressed between lexical items which are linguistically salient and which are extractable from the given lexical resources. The resources as such differ considerably in structure and content. We therefore cannot expect that the richness of the results is the same for every monolingual WordNet. The general approach is that if anticipated information is present in a given resource and if it easily extractable by semi-automatic means then it will be stored in the final multilingual database. In this respect the design will provide maximum flexibility in order to store semantic information without deviating too much from the EuroWordNet resource while keeping at the same time the monolingual networks individual and autonomous, that is without making too many commitments for building of the resources.

Furthermore, during the implementation of the project and the actual development of synsets it might turn out that some of the language internal relations and data types are not practical for the purpose of the project or will hardly occur from the lexical resources. In addition, it might turn out that new relations might need to be added or that some of the EuroWordNet relations might not need to be expressed but the idea of the functional specification is that all potential problems, aspects and relations are as much as possible anticipated.

As far as the language internal relations encoded so far as in the Greek WordNet, these are summarized in the subsequent section. Following on from this, it is proposed in a separate section that additional internal links are added in the Greek WordNet and we also point out a case that might impose the necessity for the introduction of a special case of synonymy for the case of Greek

2.3.2. Language Internal Relations Encoded So Far in Greek WordNet

SYNONYMY

Synonymy is by default the basic semantic relation that is used not only in Greek WordNet but also in every monolingual WordNet within BalkaNet network. Synonymy is a widely used semantic relation and has been used in almost all-semantic networks since the structure of synsets is based on synonymic relations. After all the advantage of the synset structure is that equivalent meanings of terms are explicitly encoded in the entry structure. A term is a synonym of another term if the former can replace the latter in any context without altering the meaning of the sentence. In EWN project a major distinction was made across synonymic relations which derived from the observation that even though terms sometimes hold the same meaning with others and thus can replace them in any context whereas others can replace them only in particular context or under particular circumstances (e.g. particular types of texts, speech etc.). Consequently, the two kinds of synonymy used in EWN are: exact-synonymy (EQ_SYNONYM) and near-synonymy (NEAR_SYNONYM). Both synonymy relations are going to be used in Greek WordNet. Synonym terms across language are

encoded as near-synonyms whereas within a monolingual WordNet the exact-synonymy relations might well be encoded. The near-synonymy relation is mostly going to be used for linking semantically close nouns (i.e., nouns close in meaning) which however share distinct hyponyms and as such cannot be encoded as identical concepts regardless of their context. On the other hand X_POS synonymy relation is going to be rather useful in linking nouns and verbs that both refer to the same event or nouns and adjectives that refer to the same situation.

HYPONYMY / HYPERNYMY

Hyponymy is a fundamental relation around which the WordNets are constructed since hyponymy along with its complementary relation (i.e. hyperonymy) link synsets as mixed conjunctive and disjunctive sets creating thus semantic chains in the lexical hierarchy. Hyperonymy and hyponymy are inverse, asymmetric and transitive relations, i.e. if Y is a kind of X, then X is HYPERONYM of Y and Y is a HYPONYM of X. An example taken from Greek language:

"*ερπετό*" has HYPERONYM "*ζώο*", ("*reptile"* has HYPERONYM "*animal"*)
"*ζώο*" has HYPONYM "*ερπετό*", ("*animal* has HYPONYM "*reptile"*)

A hyponymy relation implies that the hyperonym (the more general term) may substitute the hyponym (the more specific subtype) in a referential context but not the other way around. A referential context is a context where only the set of discourse entities is considered, whereas grammatical; register, pragmatic and other non-semantic properties of the considered words or context are neglected.

Various lexicosyntactic patterns have been reported[1] for tracing hyponymic relations such as:

- *such NP as {NP,} * {or/and} NP*
- *NP {,NP} * {,} or other NP*
- *NP {,NP} * {,} and other NP*
- *NP {,} including {NP,} * {or/and} NP*
- *NP {,} especially {NP,} * {or/and} NP*

In order to elicit the implicational relation between the hyponym and the hyperonym described above, different diagnostic tests with specific phrases can be used.

2.4. ROMANIAN WORDNET

Romanian WordNet has now a number of 6729 synsets. Romanian WordNet set is composed by an implemented set of Base Concepts plus a set of concepts chosen by two criteria which will be described bellow.

[1] For further information with respect to the hyponymic patterns please refer to M.A Hearst "Automated Discovery of WordNet Relations" pp.134

We implemented the majority of concepts proposed until now. Namely, from BC1 (Subset 1) and BC2 (Subset 2) common set which has 5000 concepts we translated into Romanian language a total of 4767 concepts. You can see a statistic in Table 2.3:

Number of Common Concepts (CC) proposed	Number of CC realized by Romanian side
BC1 1310	1230
BC2 3690	3537
TOTAL 5000	4767

Table 2.3: The Set of Common Concepts implemented in Romanian WordNet

We couldn't implement the entire common set due to the fact that some concepts have no correspondence in our language. Anyway, in the next month we will revisit the remaining concepts and try to cover some others.

The two other criteria used by Romanian side were:

1. An external criterion: The concepts covered should be translated in as many languages as possible. This could be a good criterion for choosing future concepts.
2. An internal criterion: From Romanian Explanatory Dictionary we chose a set of words with a highest occurrence frequency in our dictionary. These words have a great productivity for our language because they are used in many definitions. Then we translated those words in English and therefore we obtained a set relevant to our language.

The relations in our WordNet were inherited from original English WordNet. The relations can be seen in Table 2.4

RELATION TYPE	**NUMBER OF OCCURRENCES**
BE IN STATE	4
NEAR_ANTONYM	332
SUBEVENT	66
EQ_HYPERONYM	70
NEAR_SYNONYMYM	10
HYPERONYM	5380
CAUSES	40
HOLO_PORTION	17
HOLO_PART	294
EQ_HYPONYM	17

Table 2.4: Relations implemented in Romanian WordNet

2.5. SERBIAN WORDNET

Serbian WordNet has been implemented following the expand model and using English WordNet 1.5 as the basis. To date, 1869 synsets from Subset 2 have been translated, and the completion of the full set of 5000 synsets is targeted for the end of this year. Only two semantic relations have so far been systematically introduced in Serbian WordNet, namely those of synonymy and hyponymy/hyperonymy. Since the initial set of Base Concepts has been implemented manually, without the help of VisDic, only H/H relations have been established among the first 1305 synsets. Later, when we started to use VisDic for Subset 2 all other relations have automatically been inherited from WN 1.5. It should, however, be noted that all those relations (except hyponymy/hyperonymy) still need to be carefully re-examined.

The existing semantic relation in Serbian WordNet, for the 1869 implemented synsets are summarized in Table 2.5:

RELATION TYPE	NUMBER OF OCCURRENCES
HYPERONYM	1725
HOLO_PART	18
EQ_GENERALIZATION	326
HOLO_MEMBER	17
NEAR_ANTONYM	114
SUBEVENT	60
CAUSES	14
NEAR_SYNONYM	1
EQ_METONYM	1
HOLO_PORTION	1
BE_IN_STATE	18
EQ_DIATHESIS	2

Table 2.5: Relations implemented in Serbian WordNet

2.6. TURKISH WORDNET

From the beginning of the project, we adopted the expand model, using a subset of English WordNet 1.5. Up to now, we have translated a total of 5,000 synsets, in which the 1310 Base Concepts of the EuroWordNet Project are also included. The remaining 3,690 synsets have been jointly chosen by the partners (the selection process was described in detail in Deliverable 4.1).

We used the original XML file for WN 1.5 for the translation process. The only difference between the XML file for Turkish and that for English is the contents of the <SYNONYM> tag which contains the <LITERAL> tag for synset members and the <SENSE> tag for the corresponding sense numbers (we also added Turkish glosses for each synset, in line with the decision taken by the BalkaNet Consortium). Basing the translation work on the original

XML file for English WordNet allowed us to automatically inherit the relations of these 5000 synsets as they are encoded in WordNet 1.5.

As a result, we have obtained a subset of Turkish WordNet containing the relations shown in Table 2.6 below.

RELATION TYPE	NUMBER OF OCCURRENCES
SYNSET	5000
LITERAL	7167
HYPERONYM	4779
HOLO_PART	328
EQ_GENERALIZATION	2719
HOLO_MEMBER	127
NEAR_ANTONYM	730
SUBEVENT	293
CAUSES	63
NEAR_SYNONYM	144
EQ_METONYM	86
EQ_DIATHESIS	5
BE_IN_STATE	180
HOLO_PORTION	22

Table 2.6: Relations implemented in TWN

The automatic adoption of relations from another wordnet obviously creates a validation problem. Although we observed that the transferred relations make sense in most cases, there might be some case where they do not. Thus, each relation that has been transferred from WN 1.5 has to be manually validated. As of November 2002, we have already validated

all EQ_GENERALIZATION relations

all CAUSES relations

all EQ_DIATHESIS relations

all HOLO_PORTION relations

most HYPERONYM relations and

most BE_IN_STATE relations

All relations we have adopted from WN 1.5 have been used in the same sense as defined in EuroWordNet[2]. There are some additional remarks, however, that we should make regarding CAUSES and BE_IN_STATE relations:

2.6.1. CAUSES and IS_CAUSED_BY Relations in Turkish

[2] Vossen, P. (ed.), EuroWordNet General Document, EuroWordNet (LE2-4003, LE4-8328)

As part of the initial 5000 synsets, we have already inherited 63 CAUSES relations. A visual examination of these reveals that CAUSES has a strong relation to the following suffixes in Turkish:

The examples given below are taken from the latest version of TWN.

a) -(H)t (causative suffix)[3]
acı**t**mak (00040663-v, cause pain:1) CAUSES acımak (01211071-v,ache:4)

b) –DHr (causative suffix)
değiş**tir**mek (00072540-v,change:13) CAUSES değişmek (100064108-v, change:1)

c) -(H)l (passive suffix)
dökmek (01184040-v, pour:3) CAUSES dök**ül**mek (01182440-v, flow:10)

d) -(H)n (passive/reflective suffix)
beslemek (00670333-v, feed:6) CAUSES besle**n**mek (00670058-v, feed:5)

Turkish allows us to use these suffixes with most verbs. Thus, it is a fully automatic process to obtain these derived verbs from existing root verbs in our wordnet. However, as we mentioned above, at the first stage, we will not include all possible derivations of all Turkish verbs in our wordnet, but will only include those derived from verbs which exist and are fully described and linked in Turkish WordNet.

There are also some cases where antonymous support verbs appear in a CAUSES/IS_CAUSED_BY relation:

devr**etmek** (01266189-v,transfer:12) CAUSES devr**almak** (01266689-v,change hands:1)

rahatsız **etmek** (01035073-v, displease:1) CAUSES rahatsız **olmak** (01011964-v, dislike:3)

Most Turkish adjectives can be used as causative verbs by adding the support verb “etmek” (make) and the intransitive form of such verbs is constructed by attaching the support verb “olmak” (be) to the same adjective.

2.6.2. BE_IN_STATE and STATE_OF Relations in Turkish

We observed that BE_IN_STATE and STATE_OF relations, as defined in EuroWordNet, can be used to encode the semantic relation between:

a) adjectives and state-denoting nouns derived from them using the suffix -lHk:

özgür**lük** (08560710-n, freedom:1) BE_IN_STATE özgür (00797606-a, free:17);

[3] H represents a meta-character denoting the high vowels ‘ı, i, u, ü’. Thus each morpheme here actually stands for a set of allomorphs.

b) adjectives and state-denoting nouns derived from them using the suffix -iyet:

hürr**iyet** (08560710-n, freedom:1) BE_IN_STATE hür (00797606-a, free:17);

The suffix –iyet is of Arabic origin and is no more productive in modern Turkish. We have a limited number of lexicalizations involving this suffix which we will manually select and include in our wordnet.

c) nouns derived from nouns using the suffix -lHk:

köle**lik** (00233489-n, slavery:1) BE_IN_STATE köle (06351032-n, slave:1)

3. DISCUSSION ON THE SEMANTIC RELATIONS THAT CAN BE ENCODED

There are some relations each partner suggested in consideration of their language-specific requirements. All these relations are candidates at the moment and should be revised by the BalkaNet Consortium before they are accepted and implemented, since the relation discussed may be a common relation for all BalkaNet languages.

3.1. BULGARIAN

3.1.1. Multiple Hyperonymy Relations

Multiple hyperonyms have occasionally been encoded in WordNet. In the English database WordNet 1.5 only 582 synsets are in relation with two hyperonyms.

Some examples:

ILI: 02199263-n	(brick:1)	ILI: 02376836-n	(cricket ball:1)
H1 : 02273139-n	(ceramic:1)	H1 : 02103632-n	(ball:2)
H2 : 08885624-n	(building material:1)	H2 : 02377053-n	(cricket equipment:1)

We assumed that in every hyperonymy / hyponymy relation n hyperonyms can appear, where $n \geq 1$. If $n = 1$, the semantic meaning of the hyponym is a proper subset of the semantic meaning of the hyperonym. Conjunction and disjunction are applied in WordNet to the hyperonymy relation:

A spoon is both a container and a piece of cutlery.
A knife is either a weapon or a piece of cutlery.

If $n = 2$ (number of hyperonyms to a given synset), we presume empty or not empty intersection of both sets of semantic meanings. The lack of the intersection between the sets of semantic meanings of two hyperonyms H1 and H2 is inherited by its immediate hyponym H0 which means that H0 is a subset either of H1 or H2 (disjunction is applied). If the sets of semantic meanings of the hyperonyms H1 and H2 have an intersection, the set of semantic

meaning of their common immediate hyponym H0 is equal to the intersection or its subset. Thus the hyponym H0 inherits a semantic meaning both from H1 and H2 (conjunction is applied) and thus from the higher hyperonyms.

Some examples:

ILI: 08817229-n (O:2, atomic number 8:1, oxygen:1)
the most abundant element
H1 : 08805286-n (chemical element:1, element:6)
any of the more than 100 known substances that cannot be separated into simpler substances
H2 : 08938440-n (gas:5)
a fluid in the gaseous state having neither independent shape nor volume

ILI: 02562976-n (стъклена чаша:1) drinking glass:1, glass:2
H1 : 01990006-n (контейнер:1, резервоар:1) container:1
H2 : 02563503-n (стъклария:1, стъкларски изделия:1) glassware:1, glasswork:1

ILI: 02391765-n (завеса:1, перде:1) curtain:1, drape:1, drapery:1, mantle:2, pall:1
H1 : 02043015-n (покъщнина:1, обзавеждане:1) furnishings:2
H2 : 03071953-n (покривало:2, параван:1) blind:2, screen:7

Some disadvantages arise because the multiple hyperonymy relations are not consistently encoded in the English database and this would cause a lot of discrepancies between WordNet structures. We should try to encode multiple hyperonyms in different languages and to keep correspondence with English structure. The first candidate from the multiple hyperonymy group has to be equal with the English one, if in English structure is encoded only one hyperonymy. The results of such approach should be avoiding of some artificial hierarchy between words, however the correspondence with the WordNet structure would remain.

3.2. CZECH

When working both with the noun and verb synsets in Czech we have faced the problem of the translation equivalents and corresponding gaps with regard to English. There are two kinds of cases where it is not possible to find the eq_synonyms (or even near_synonyms):

1. The Czech synsets do not have corresponding counterparts in WN 1.5 due to differences in lexicalizations and conceptualizations between Czech and English:

a) Czech synsets do not have eq_synonyms in English at all,
b) Czech synsets do not have eq_synonyms in WN 1.5 but we have been able to find their English equivalents in general.

A typical example: in Czech there is a synset represented by the literal *náledí* which does not have the eq_synonym in ILI but its regular translation equivalent in British English is *black ice* (see e.g. NODE, 1998, p.182). There are numerous examples of this sort – the differences between WN 1.5 (American English) and BrE are noticeable. We may offer more examples of this sort, our present list of such Czech nouns contains about 1200 items and list of verbs – about 500 items but these numbers are far from final.

2. The second collection of Czech items without eq_synonyms in ILI comprises Czech items that are typologically different due to the highly inflectional nature of Czech with its rich formal and derivational morphology. Because of that in some typical cases the straightforward English translation equivalents either cannot be easily found or have to be substituted by the various syntactic constructions or context dependent equivalents have to be searched for. At the present moment at least four types of basically morphological phenomena causing the gaps should be mentioned:

1) verb aspect: imperfectives, perfectives, iteratives
2) reflexive verbs
3) verb prefixation (single, double)
4) deminutives (noun derivation by suffixation)
5) move in gender (noun derivation by suffixation)
6) other types of word form derivation (word derivation nests, families).

The mentioned phenomena are, in our view, relevant in the WordNet context and they can be generalized in the following way:

3.2.1. Aspect Opposition

Actions and activities in Czech can be seen as (in fact) a ternary relation: imperfective – iterative – perfective,
where **imperfective** verbs express actions and activities **unbounded in time** (*číst – read*),
perfective verbs denote actions and activities bounded in time (*přečíst – read through, to read completely*) and
iteratives refer to the regularly repeated actions and activities bounded in time, i.e. primarily they are classified as imperfectives (*čítávat – read regularly, repeatedly*).

The question to be answered is: shall we have one synset for each of the mentioned verb types or should we keep the information about them in one synset for all? In our opinion, the appropriate solution, at least within the Czech WordNet, would be to introduce appropriate internal language relation(s) that could link together the respective synsets. In the case of the aspect we would suggest to introduce internally in Czech WordNet as a new kind of ILR something like X_HAS_IMPF, X_HAS_PERF, X_HAS_ITER attributes. With regard to ILI, however, it seems to be reasonable not to project (since it would call for a rather detailed verb subclassification) these differences further. At least this standpoint was taken during the development of EWN 2 when these issues were discussed. However, for BalkaNet it may appear reasonable to reconsider these questions once more because more typologically different languages are being involved, i.e. Greek, Turkish, plus three Slavonic languages and Romanian.

3.2.2. Reflexives

At least the three relevant types of the reflexive verbs in Czech have to be taken into consideration:

- reflexiva tantum (full grammatical reflexives), e.g. *smát se* (*laugh*)
- verbs expressing reflexivity, e.g. *holit se* (*to shave himself*)
- verbs expressing reciprocity, e.g. *milovat se* (*to love each other*)

The suggestion is to keep them systematically as separate synsets – as it is standard in Czech dictionaries but again as we have pointed out above – we rather would not project them into ILI and keep the appropriate ILR attributes in Czech WordNet only. Other cases of reflexive verbs and their meanings belong mainly to the syntactic level. For BalkaNet languages we propose to compare all the types of reflexives and if it appears that reflexivity is a relevant category a solution has to be looked for.

3.2.3. Prefixed verbs

They combine the aspect distinctions with iterativity and various types of the distributed actions (*posbírat – pick one after another*), thus they are sources of many gaps. Some of them can be translated by (English) phrasal verbs but there is not very much regularity in this respect. The use of the relation HAS_NEAR_SYNONYM appears as a possible solution, though within Czech WordNet we intend to mark these cases completely.

3.2.4. Gender pairs

They display binary semantic opposition – male: female, and the question is again similar as above: shall we have this distinction in one synset as it is in English WordNet or it is reasonable to keep them apart as a separate synsets and have a special attribute with two values for them, say: X_HAS_MALE – X_HAS_FEMALE? The solution adopted in WordNet 1.5 is acceptable also in Czech (Klímová, Pala, 2000), however I suspect that the situation in Bulgarian, Serbian and Croatian will be quite similar to that in Czech.

3.2.5. Diminutives

They display a sort of ternary semantic opposition as in the case of aspect, however, there is, a relevant difference: one of the attributes expresses an emotional attitude of the speaker in a lexicalized way. Thus, as we have indicated above, the following cases can be found with Czech diminutives:
- standard, as in *dům, domek* (in English *house, cottage*)
- small thing as in *domek, domeček* (*small house, Wendy house*)
- emotional attitude as in *domek*, *domeček* (something like *my dear little house*).

To preserve this information in Czech WordNet we suggest to introduce (tentatively) the following attributes: X_IS_SMALL and X_EXPRESSES_POSITIVE_EMOTION.

Generally, however, we are prone to the opinion that it would be reasonable to re-evaluate the ILR table and to consider the possibility of adding new attributes to the ILR with consequences for ILI as well: when for other typologically different languages the WordNet databases are going built, such as Greek, Turkish or other languages – Romanian, Bulgarian, Serbian and Croatian) it will be necessary to look for a more consequent and complete solution anyway. Our suggestion (based on the experience with GENELEX project which included French, Czech, Polish and Hungarian) would be to have the optional attributes in the ILR table (or slots specific to the particular languages). What remains to be answered is the question how to reflect these language specific attributes in ILI and how to relate them reasonably but the idea of optional slots should certainly be considered.[4]

[4] Pala K., Klímová J., Application of WordNet ILR in Czech Word-formation, Proceedings of LREC, Athens, June 2000, p.987-982.

3.3. GREEK

In the following paragraphs we summarize which of the language internal relations that are currently missing from the Greek WordNet need to be incorporated. During selection process we pay particular attention to the POS in which a terms might belong, enabling at the same time X_POS relations in cases where necessary.

Language internal relations that will be encoded in Greek WordNet are:

3.3.1. Belongs to

Each synset is going to have a BELONG_TO link to each of the domain labels of the BalkaNet semantic network in order to indicate the specific domain or sub-domain to which it conceptually belongs. Thus, each variant apart from a gloss and an optional sage label that will or might have attached it is also going to be linked to one or more of the conceptual domains in order to facilitate navigation of the end user not only across different synsets but also across different concepts. In particular, the user will be able to view not only terms that are conceptual equivalents, i.e. terms that belong to the same synset but also he/she will be able to view terms that belong conceptually to the same domain(s). The latter is rather important for the performance of conceptual indexing or classification tasks.

3.3.2. Antonymy

The relation of antonymy refers to the semantic opposition between concepts. Even though it is not a fundamental relation between nouns in comparison the hyperonymy/hyponymy and synonymy relations, nevertheless is plays a crucial role for the project's final application and for the accurate representation of lexical information in a semantic network. One problem encountered towards encoding antonymic relations is WordNet concerns the limited amount of information encoded within various lexical resources as far as semantic opposition is concerned. This is why antonymy might be encoded in Greek WordNet to a limited extend and in cases were an oppositional established link can be verified against lexical resources. However, antonyms pose some difficulties during encoding which should be taken into account. In particular, noun antonyms might have the same immediate hyperonym, which might difficulties during encoding the BELONG_TO semantic relation. Finally, since antonymy is a relation found mostly between descriptive adjectives and taken into account the limited distribution of adjectives within each WordNet, the restricted usage of antonymy might also be accounted.

3.3.3. Meronymy / Holonymy

Meronymy is the relation between an object and its constituents or proper parts. Such a relation is usually between a noun that denotes the whole and a noun that denotes its parts, i.e., it is a noun-noun relation. This relation has an inverse: if S is a MERONYM of B, then B is said to be a HOLONYM of S. meronymy and holonymy relations stand mostly for concrete objects, that is mostly for 1st and perhaps for some 2nd order entities. Such relations do not

seem necessary for terms denoting quantities, where small units of measurement are parts of larger units at every level in the hierarchy. Even though several types of meronymic relations have been proposed to the literature the one that will be encoded within the framework of the project is the IS_A_COMPONENT_OF relation.

3.3.4. Derived_from

Besides strictly semantic relations that hold between terms some other kinds of semantic links might well occur. Such links are found within terms that have a primitive morphological connection, which is also applicable to the level of semantics. More specifically, word forms produced by a common stem inherit not only stem's morphosyntactic features but might also inherit the stem's semantics. Such cases will be encoded in the Greek WordNet via the DERIVED_FROM relation, where the latter will be restricted to the semantic properties of the terms in question. Moreover, DERIVED_FROM is going to form a rather useful relation while encoding adverbs, since the majority of the latter is derived from adjectives. Taking into account that adverbs are rather straightforwardly organized in a semantic network, i.e., synonymy and sometimes antonymy are the only relations usually recognized, the main relation used for linking adverbs with other sysnets of the network is the DERIVED_FROM relation in which the adverb inherits the sense of the base adjective to which it is linked to[5].

3.3.5. Role/Agent, Involved Agent, Role/Instrument, Involved Patient

The language internal relations discussed so far can be expressed either between pairs of 1st, 2nd or pairs of 3rdOrderEntities respectively, but never across these types. In order to account for across ontological types relations we suggest that semantic links of the *roles* and *functions* types are represented between concepts of the network. In particular, if the relation goes from a concrete or mental entity (only nouns denoting 1st or 3rdOrderEntities) to verbs or event denoting nouns (2ndOrderEntities), it will be called *role*, the inverse from events (2ndOrderEntities) to concrete or mental entities (nouns) is called *involved*.

3.4. ROMANIAN

We analyze the possibility to encode other relations beside those existing in English WordNet. After deciding which will be the set of new relations we will use automatic techniques for extracting from Romanian Explanatory Dictionary those words which will be in specified relations. After every partner will finalize the list of relations it will be better to have a discussion about possible common relations.

3.5. SERBIAN

It is our opinion that it would be most useful, at least for Serbian, to consider possible derivational relations in addition to semantic relations that have already been recognized by WordNet. These relations would be particularly valuable in cases of the so-called structural derivation, where the meaning of a derived word can be predicted from the original word and

[5] Each adverb is linked to a particular adjective sense and not to an adjective base form.

the derivational process. These cases could be easily extracted from an explanatory dictionary, as derived words have stereotypical definitions. For instance, pecyenxe in Serbian (Eng. (a) fry) is defined simply as a verbal noun of the verb pecxi (Eng. (to) fry) and the same is true for all verbal nouns in the Serbian explanatory dictionary. A similar pattern is followed by possessive adjectives: vojvodin (Eng. Duke's) is defined as "koji pripada vojvodi" (Eng. "which belongs to a Duke"). However, contrary to verbal nouns that exist in the dictionary for all listed verbs, only for some of the nouns corresponding possessive adjectives can be found as dictionary lemmas (although many other noun lemmas can have possessive adjectives as well).

We believe that derivational phenomena deserve a thorough investigation since it seems that their inclusion in WordNet could have some very valuable effects. Of course, we have to bear in mind the fact that relations of this type could be established between synsets only when the same derivational process applies to all members of a synset

3.6. TURKISH

Our basic approach is to use the already defined EuroWordNet relations, especially the implemented ones, as much as possible. There are 84 relations and reverse relations defined in EuroWordNet but most of them were not implemented in any of the individual wordnets of the project.

There are several other (morpho-)semantic relations that could be added to Turkish WordNet. Especially in the case of verbs, the morphology is highly complex, fully productive and predictable. A verb root may take several suffixes following each other. The morphology of nouns and adjectives is also highly complex.

However, as a general rule, we are not going to systematically add all complex derivational forms at the first stage. Although adding all derivational forms would provide valuable information on the Turkish language, linking the derived forms to their base forms, linking them to the ILI, writing Turkish glosses for each and encoding the semantic relations between them is a very time-consuming task which goes beyond the objectives of the present project.

Although we do not plan to do it systematically, we have already included some complex derived verbs as a result of the expand model we used for the first 5000 synsets. For instance, the complex verb "*sertleştirmek*" (sert + leş + tir + mek) (00247885-v, make stiff:1, make stiffer:1, stiffen:1) is a member of the first 5000 synsets since the English verb "stiffen" cannot be translated into Turkish in any other way.

3.6.1. Word-to-Word Relations vs. Synset-to-Synset Relations

A basic problem of the current XML structure of the Balkanet WordNet is that it does not allow us to encode relations between lexical items rather than between synsets. These relations can be easily and safely extracted from our existing resources.

We think that encoding a word-to-word relation as a synset-synset relation is not meaningful. This results in a loss of information, which can be valuable for NLP tasks. For instance, consider we establish a CAUSES relation between the following synsets:

{**öldürmek**; **gebertmek**; temizlemek} (00758542-v, kill:5)

{nalları dikmek; hakkın rahmetine kavuşmak; **ölmek**; **gebermek**; mevta olmak}

(00216283-v, die:6)

we lose the valuable information that the causation relation in grammatical terms is "öldürmek" (öl +dür + mek) of the first synset and "ölmek" of the second synset, and "gebertmek" (geber + t + mek) of the first synset and "gebermek" of the second synset. On the other hand, the verb "temizlemek" is not such a morphological causative form of a verb. This example shows that the semantic relation is not morphologically applicable to all members of the synset.

There are many other cases where we want to include purely morphological relations, which obviously exist between lexical items and not synsets. For instance, the suffix –cH has at least four systematic semantic effects and several non-systematic semantic effects. The first four examples in Table 3.1 show the different predictable semantic usage of the –cH suffix and the rest are examples for cases where the semantic relation cannot be generalized.

WORD FORM	DERIVED WORD	RELATION
balık (fish)	Balık+çı (fisherman)	ITEM SOLD/PRODUCED and PRODUCER / SELLER
dedikodu (gossip)	dedikodu+cu (gossiper)	ACTION and PERSON WHO HABITUALLY DOES IT
pilav (rice)	Pilav+cı (rice lover)	THING and LOVER OF THING
Mao (Mao)	Mao+cu (adherent of Mao)	PERSON and ADHERENT
kira (rent)	kira+cı (lessee)	NOT CLEAR
yol (road)	yol+cu (passenger)	NOT CLEAR
dava (lawsuit)	Dava+cı (plaintiff)	NOT CLEAR

Table 3.1: Examples of words derived by the –cH suffix.

Establishing links between individual lexical items has been attempted in EuroWordNet by the DERIVED and PERTAINS_TO relations, but these relations have never been implemented[6].

In order to encode relations between individual lexical items, we have to modify our existing XML structure. This could simply be achieved by creating a new file containing synset members (literals) and sense numbers and assigning a unique ID to each unique literal-sense combination (to be referred to as, say, LIT_ID in the XML file). The structure of a line in our XML file will change as follows:

[6] see Vossen, P. (ed.), EuroWordNet General Document, EuroWordNet (LE2-4003, LE4-8328), p. 37)

<SYNSET><ILI>00003702-a</ILI><POS>a</POS><SYNONYM><LITERAL>aborning<SENSE>1</SENSE></LITERAL> …

<SYNSET><ILI>00003702-a</ILI><POS>a</POS><SYNONYM><LIT_ID>0178455</LIT_ID> …

The following section discusses several (morpho-)semantic relations we plan to include in Turkish WordNet. Since these relations are morphosemantic, each time we set a morphologic relation between two word forms of synsets, we form a semantic relation between the synsets at the same time. Therefore the restriction of setting the relations only between synsets does not cause implementation problems for the near future.

3.6.1. INVOLVED_X Relations and CO_X_Y Relations

According to EuroWordNet, INVOLVED_X relations are between 2nd order entities on the one hand, and AGENTS, PATIENTS, RESULTS, INSTRUMENTS, SOURCE_DIRECTIONS, TARGET_DIRECTIONS and LOCATIONS on the other hand. In other words, INVOLVED_X relations link the following pairs:

EVENT OR ACTION - AGENT
EVENT OR ACTION - PATIENT
EVENT OR ACTION - RESULT
EVENT OR ACTION - INSTRUMENT
EVENT OR ACTION - SOURCE_DIRECTION
EVENT OR ACTION - TARGET_DIRECTION
EVENT OR ACTION - LOCATION

CO_X_Y relations, on the other hand, link AGENTS, PATIENTS, RESULTS and INSTRUMENTS to each other, but never explicitly involve the EVENT OR ACTION itself. In other words, they link the following pairs:

AGENT - PATIENT
AGENT - RESULT
AGENT - INSTRUMENT
PATIENT - RESULT
PATIENT - INSTRUMENT
RESULT - INSTRUMENT

As native speakers of Turkish, a language greatly influenced by Arabic and Persian until the mid-20th century, we retained some knowledge of so-called “consonant radicals” in Arabic. In thousands of Arabic words which still survive in Turkish, we have a root formed by three consonants and around 40 frames for inflecting these roots, which then acquire certain predictable meanings.

For example, using the 3-letter root K-T-B and the respective frames for EVENT, AGENT, PATIENT, LOCATION, etc., we obtain:

IKTAB (to write, ACTION)
MEKTEB (school, LOCATION, where you write)
KATIB (scribe, AGENT, one who writes)
KITAB (book, RESULT, written thing)
MEKTUB (letter, RESULT, written thing)

As can be seen, the morpho-semantic system of Arabic is strikingly similar to the EVENT, AGENT, PATIENT, INSTRUMENT, RESULT, LOCATION, and DIRECTION paradigm and is fully predictable. There are several additional inflection frames in Arabic which go beyond basic concepts such as AGENT-PATIENT etc. and thus constitute a very neat structure of semantic relations between verbs[7].

Although a large number of lexicalizations following the rules of Arabic derivational morphology exist in modern Turkish, the system is not productive today. So, there is only a limited number of lexicalized forms in the language and a native speaker of Turkish does not generate new words using these rules. For this reason, we are not going to undertake automatic generation and automatic relation building for these lexical items. For instance, we will not automatically generate the action-form “iktab” simply because we have the object-form “kitab” in our wordnet, but only include those forms which actually exist in our language.

In this context, we feel that the distinction between INVOLVED_X and CO_X_Y relations is not necessary. In addition, we have to drop the "CO_" part of the relations since we do not always deal with compounds in our language. As a result, we initially planned to define the following 21 new relations that would comply with our conceptualization of the issue:

EVENT_AGENT
EVENT_PATIENT
EVENT_INSTRUMENT
EVENT_RESULT
EVENT_DIRECTION
EVENT_LOCATION
AGENT_PATIENT
AGENT_INSTRUMENT
AGENT_RESULT
AGENT_DIRECTION
AGENT_LOCATION
PATIENT_INSTRUMENT
PATIENT_RESULT
PATIENT_DIRECTION
PATIENT_LOCATION
INSTRUMENT_RESULT
INSTRUMENT_DIRECTION
INSTRUMENT_LOCATION
RESULT_DIRECTION
RESULT_LOCATION

[7] http://www.elsnet.org/arabic2001/black.pdf

DIRECTION_LOCATION

But this would cause additional confusion regarding relation tags. As a result, we decided to maintain the relation names of EWN but explain in our documentation that a CO_X_Y is not necessarily used for compounds in Turkish WordNet. If we want to add a new EVENT_X relation, we will add it as an INVOLVED_X relation and if we want to add a new X_Y relation which doesn't involve EVENT, we will add it as a CO_X_Y relation.

3.6.2. Verbs Derived From Nouns and Adjectives Using The Support Verbs "Etmek" and "Olmak"

Example: hareket (movement) – hareket **etmek** (to move)
deli (mad) – deli **olmak** (to go mad)

We could define an XPOS_NEAR_SYNONYM or an INVOLVED_RESULT relation between the two sides.

3.6.3. FUZZYNYM and XPOS_FUZZYNYM

These are two non-specified relations which could be used to mark synsets you wish to relate to each other but do not know which relation to use. It is possible to use these relations as a temporary repository of related pairs.

3.6.4. "Become Verbs" Derived From Nouns and Adjectives Using the Suffix "-laşmak"

Examples: a) taş (stone) → taş+laş+mak (to petrify)
b) genç (young) → genç+leş+mek (to become younger)

The relation between the root noun and the derived verb in Example (a) above can be defined as an INVOLVED_RESULT relation.

3.6.5. Verbs Derived from Nouns and Adjectives Using the Suffix "-lamak"

Examples: a) tuz (salt) → tuz+la+mak (to add salt)
b) yavaş (slow) → yavaş+la+mak (to slow down)

The relation between the root noun and the derived verb in Example (a) above can be defined as an INVOLVED_INSTRUMENT relation.

3.6.6. Verbs Derived From Nouns Using the Suffix "-lanmak"

Examples: renk (colour) → renk+len+mek (to acquire colour)

If we add this suffix to a noun X, the derived verb has the meaning "to acquire X". There is no corresponding relation in WN 1.5 or EuroWordNet..

3.6.7. Reflexive Verbs Derived From Action Verbs Using the Suffix "-(A)Nmak"

The suffix (A)nmAk is not a productive suffix but survives in a limited number of reflexive verbs. There is no corresponding relation in WN 1.5 or EuroWordNet. For these verb pairs, we could manually add a relation which could be entitled "HAS_REFLEXIVE_VERB".

Examples: yıkamak (to wash) → yıka+ n +mak (to wash oneself)
süslemek (to decorate) → süsle+ n +mek ("to decorate oneself", to make up)
taramak (to comb) → tara+ n +mak (to comb one's hair)

3.6.8. Classes and Instances

Example: İstanbul – {şehir, kent} (city)
Özlem – {isim, ad, özel isim}(proper name)

Classes and instances could be linked via the BELONGS_TO_CLASS and HAS_INSTANCE relations defined in EuroWordNet[8].

4. TECHNIQUES USED FOR EXTRACTING RELATIONS

Most of the partners used semi-automatic or automatic techniques for extracting relations. These techniques are described in detail under each language's section, providing quantitative data and examples where possible.

4.1. BULGARIAN

4.1.1. Automatic Translation

Automatically assignment of Bulgarian translations to the English synsets is necessary step in the work. Automatic translation of English literals with the English - Bulgarian electronic dictionary is applied in two different ways: automatic translation of the first English literal from the corresponding synset and automatic translation of every English literal from the corresponding synset. A lot of senses are assigned to one word (some of them not correct) and a professional translator should decide what the real meaning correspondence was and eliminate unwanted words.

[8] Vossen, P. (ed.), EuroWordNet General Document, EuroWordNet (LE2-4003, LE4-8328), p. 37)

4.1.2. Automatic Assignment of Additional Synonyms

Automatic assignment of synonyms from the electronic synonymous dictionary is used as follows: assignment of synonyms to the first Bulgarian literal from the synset and to every Bulgarian literal combined with a unification of the candidates. As a result a lot of candidates appeared and some of them were not in semantic relation of equivalence. For example hyperonyms or hyponyms are included in one lexical entry.

4.1.3. Automatic Assignment of Hyperonyms

The English hyperonyms which are members of Bulgarian BC (Subset 2) were assigned automatically to Bulgarian synsets (at the same time with previous stages). In spite of the fact that the hyperonymy relations may be different in two languages automatically assignment of the English hyperonyms helps a lot.

4.2. GREEK

Various techniques have been employed during selection and extraction of the semantic relations that exist among synsets of the Greek WordNet, ranging from automatic to completely manual techniques. More specifically, the available lexical resources are mainly limited to explanatory dictionaries in electronic forms and a small-range Greek corpus (ECI corpus) compiled by the UOA team. The methodology we followed while processing the abovementioned resources towards encoding semantic links are summarized below:

4.2.1. Extraction of Language Internal Relations for Greek

Some of the entries encoded in the Greek explanatory dictionaries provide along with definitions of the underlying terms some semantically similar words, encoded under the relation of synonymy. In such cases synonyms of such terms were extracted automatically from the resources, making extensive usage of tools developed by the university of Athens (UOA). The same holds for the antoynymy relation with the only difference that the latter is encoded in Greek dictionaries in a much more a limited extend. A completely automated process was adopted in cases where synonyms or antonyms were provided under the respective glosses of the terms. However in quite a few cases, synonyms and/or antonyms of some terms appear at the end of the lexicographic entry of the term in question and do not provide any information on whether they apply (refer) to one or more of its definitions nor does it indicate the definition to which it refers to. In such cases semi-automatic methods were employed in order to trace the sense to which each synonym corresponds. More specifically, once synonyms of dictionary entries were automatically extracted, linguists manually processed them in order to map the latter against the correct gloss(es) encoded within WordNet synsets. The same happened in the case of hyperonymy, which in some instances was automatically defined whereas in others manual work had to take place. Particularly, much information for discovering hyponymic/ hyperonymic relations among terms is being encoded in traditional explanatory dictionaries we had at out disposal. In such resources words are defined in terms of other words, reflecting the way in which its meanings are specified. As such, definitions accompanying terms in dictionaries were processed in order to discover terms holding a semantic relation with the head entry word. However, due to the fact

that Modern Greek is a highly inflectional languages some problems arose while looking for hyponyms/hyperonyms in dictionary definitions. More specifically, many terms within a definition were in an inflected form and as such could not be automatically checked against other resources to verify the correctness of the relation. Such cases need to be tackled by a lemmatizer, which will induce all inflected word forms to their respective root forms automatically. A lemmatizer to be used in this direction is currently under development by the UOA contractor, whereas both CTI and DBLAB team plan to make extensive usage of a normalizer they have developed for Modern Greek. A normalizer performs a similar task to lemmatization with the additional feature that it assigns POS labels to the terms in question and attempts a POS resolution where necessary with quite remarkable outcomes.

Finally, with regard to the remaining lexical relations that had to be encoded in the Greek WordNet manual processing of the resources will take place to a large extend due to the limited amount of resources we have currently at our disposal. Validity of the language internal relations will be verified towards the ECI corpus.

4.2.2. Problems Encountered during Selection Process of the Greek Base Concepts

Regarding the synonymy relation there is a number of words conventionally referred to as synonyms in Modern Greek dictionaries, which nevertheless seem to need further checking before being used as such in the Greek WordNet.

Standard Modern Greek permits a good deal of vocabulary variation due to the former linguistic situation of bilingualism[9] (diglossia). Word selection varies in such cases according to whether a term is used in spoken or written context or on the types of documents it is found in (i.e. literature texts still tend to make extensive usage of word forms that do not have all their inflections expressed according to the Modern Greek inflectional system).

The problem has emerged when we possessed lexical resources form tracing the Greek Base Concepts where we encountered the phenomenon that terms listed in dictionaries as exact (direct) synonyms seemed to be related with another kind of synonymy. According to Miller et al. “two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value of it”. But what happens when this is a matter of using different registers or of speaking/writing in different pragmatic situations?

On top of that there are a few cases in Greek where terms are widely known and used for their derivational function. In such cases deciding on whether they should be included in a semantic network involves extensive usage of various corpora covering various fields of speech (both spoken and written). However, different kinds of corpora might provide us with different kinds of results, which actually reflect the present state of the language.
At this early stage of the project what we suggest is that the monolingual WordNet of Standard Modern Greek includes, alongside the range of Modern Greek literals elements productive in the present-day language, spoken or written even if they have slightly differentiated inflectional forms from the valid ones. Nevertheless, the kind of synonymy-relation that should link this kind of words is being denoted so far as synonymy* in order to be differentiated from the synonymy relations used in the EWN. We expect future research to

[9] The phenomenon of bilingualism refers to two distinct dialects used for quite a long period in Greece which were both widely used with the only differentiation that one dialect used many forms of the Ancient Greek and was through as a more formal one whereas the second one used more forms of Modern Greek and was mostly used in spoken speech.

shed light on this matter and once various tests on the resources and the terminology extracted are performed we will be able to report more concrete results.

4.3. SERBIAN

The relations that MATF will try to extract (semi-)automatically from the e-version of a Serbian explanatory dictionary are antonymy and meronymy/holonymy. However, the first tests that we have performed show that these relations have not been implemented in the dictionary systematically. For instance, the adjective svetao (Eng. bright) is defined in sense 2a as "koji nije taman, koji je otvorene boje..." (Eng. “not dark, of a light color”), that is as an opposite of its antonym adjective taman (Eng. dark). Surprisingly enough, taman does not at all refer to the adjective svetao in any of its 4 senses or 8 subsenses. Similarly, the adjective dobar (Eng. good) is in sense 1a defined explicitly as an antonym of zao, rdxav, losx, while at the same time zao (Eng. bad) does not refer, neither explicitly nor implicitly, to dobar. The same is true for definitions of losx and rdxav, synonyms of zao.

As for the meronymy/holonymy relations, we have encountered similar types of inconsistencies. For instance, cvet (Eng. flower) is defined in sense 1 as "reproduktivni organ u bilxaka koji..." (Eng. “a reproductive organ of plants that...”) while the definition of latica (Eng. petal) reads as "listicx cvetne krunice" (Eng. “a small leaf of a flower’s crown’”). On the other hand, stablo (Eng. trunk) and koren (Eng. root) have more consistent and explicit definitions: stablo is defined in sense 1a as "nadzemni deo drveta..." (Eng. “part of a tree above ground...”), and koren in sense 1 as "podzemni deo bilxke..." (Eng. “underground part of a plant”). An interesting example is volan (Eng. steering wheel), defined as "upravlxacy, kormilo automobila, bicikla, aviona" (Eng. “steering gear, rudder of a car, bicycle, airplane”). This definition is not explicit, but rather uses a genitive construction and refers to some vehicles that have steering wheels as their part, but also to some that do not.

In spite of all the problems listed above, it seems that a thorough analysis of constructions used in definitions could lead to a (semi-)automatic extraction of some useful relations from the dictionary.

4.4. TURKISH

4.4.1. Synset Extraction from TDK (Turkish Monolingual Dictionary)

Our electronic monolingual dictionary TDK has many instances of the following format:

$$\text{hw: w (, w}_i\text{)}^*$$

where, hw is the headword and w is a single word.

This format gives us the chance to extract candidate synsets out of our machine-readable dictionary, automatically. For example, the entry

abartı: abartma, mübalağa (exaggeration)

can construct the synset

abartı, abartma, mübalağa

If we generalize this example as a formula, we can write

$$hw, w_1 , w_2 , w_3 , ...w_n$$

Almost llK such forms have been extracted by using Perl scripts to parse dictionary entries. The technique and the scripts were explained in Deliverable D.3.1 in detail. The exact numbers of extracted synsets are given in Table 4.1.

Synset type	Quantity
1+1	6642
1+2	3278
1+3	974
1+4	199
1+5	30
1+6	3
Total	11126

Table 4.1: Number of synsets automatically extracted from TDK

There are also entry patterns like

$$hw: (w_i)^*, w$$

which means the headword is defined and a synonym is added at the end of the definition, separated by a comma. From such patterns we can obtain the synset

hw, w

For example the entry

endüstriyel: Endüstri ile ilgili, sınai

has the synset

endüstriyel, sınai (industrial)

All such forms have been extracted by a Perl script and the total number is 10846.

4.4.2. Extraction of Other Relations from TDK

Our monolingual dictionary enables us extracting instances of some already defined relations by using the patterns in the definitions of headword entries. By using Perl scripts we have constructed various files containing the possible relation pairs describe below. These potential pairs should be revised manually and then will be added to TWN as a future work.

HYPERONYMY

The pattern “bir tür” (a kind of) in the gloss indicates a hyperonymy relation between the headword and the word form succeeding the pattern.

anofel: Sıtma mikrobunu aşılayan bir tür sivri sinek

anofel HYPERONYM sivrisinek

There are 454 instances of “bir tür” in TDK

The pattern “bir çeşit” (a kind of) in the gloss also indicates a hyperonymy relation between the headword and the word form after the pattern.

çizme: Koncu diz kapaklarına kadar çıkan bir çeşit ayakkabı

çizme HYPERONYM ayakkabı

There are 171 instances of “bir çeşit” in TDK

More than one instances of hyperonymy relation can be established if the gloss contains the pattern “genel adı” (general name).

erdem: Ahlakın övdüğü iyilikçilik, alçak gönüllülük, yiğitlik, doğruluk gibi niteliklerin genel adı

iyilikçilik	HYPERONYM	erdem
alçak gönüllülük	HYPERONYM	erdem
yiğitlik	HYPERONYM	erdem
doğruluk	HYPERONYM	erdem

There are 81 instances of “genel adı” in TDK

The suffix “-giller” is usually used for constructing biological taxonomy terms. Therefore definitions of animals and plants usually contain this pattern, which is appropriate for extracting hyperonymy relation.

Limon: Turunçgillerden, 3, 5 m yükseklikte, kışın yapraklarını dökmeyen, beyaz çiçekli bir ağaç (Citrus limonum)

Limon HYPERONYM turunçgiller

There are 889 instances of “giller” in TDK

ANTONYMY

The pattern “karşıtı” opposite of in the gloss can be used for extracting automatic antonymy relations between the headword and the word form preceding the pattern “karşıtı”.

yumuşak: Kolaylıkla bükülen, sert karşıtı

yumuşak ANTONYM sert

There are 235 instances of “karşıtı” in TDK

5. CONCLUSION AND FUTURE WORK

As part of the future work required for this work package, each partner is going to review the relations implemented and proposed by each partner and compare it to its own set of relations. After each partner has sent comments regarding the relations described in this deliverable, the Consortium will have to make a common decision regarding the inclusion and exclusion of individual relations, the names to be attached to each of them and the respective description of their nature. Following this joint decision, each partner will implement the relations, using the extraction techniques described here.