Bulgarian WordNet - structure and validation

Svetla Koeva – Bulgarian Academy of Sciences
Tinko Tinchev*– Sofia University
Stoyan Mihov†– Bulgarian Academy of Sciences

Abstract

This paper presents a brief description of the semantic relations included in the Bulgarian WordNet. A complete and decidable formal logic for the WordNet structure is also proposed. This logic provides sufficient expressive power for all important verifications, queries, and consistency and completeness proofs required for WordNet applications. Some parameters concerning Bulgarian synsets and language-internal relations, as well as the distinctive features characterizing the completeness and consistency of the Bulgarian WordNet are laid out below.

1 Introduction

The major part of the relations encoded in the Bulgarian WordNet (BulNet) are semantic relations: ALSO SEE, CAUSE, HYPERNYMY, MEMBER MERONYMY, NEAR ANTONYMY, PART MERONYMY, PORTION MERONYMY, SIMILAR TO, SUBEVENT, VERB GROUP. There are also some morpho-semantic relations: BE IN STATE, BG DERIVATIVE, some morphological (derivational) relations: DERIVED, PARTICLE, and some extralinguistic ones: REGION DOMAIN, USAGE DOMAIN, CATEGORY DOMAIN [4]. The specification of the relations was made according to both the available descriptions of the relations [3], [5], [9] and the Bulgarian linguistic tradition. The formal representation of the relations helps us to formulate the descriptive logic for WordNet which, on the other hand, is very powerful in the applications for validation of the WordNet completeness and consistency. Our methodology defined for WordNet

*Co-author of the section WordNet Logic.
†Co-author of the section WordNet Logic.

validation was implemented in several tools and has been resulted in the very good parameters that characterize the Bulgarian WordNet.

2 Language-internal relations

The list of semantic relations presented in BulNet is based mostly on the Princeton WordNet lexical and conceptual relations, and the EuroWordNet language-internal relations. There are some differences in the treatment of some of the relations due to language specific features and linguistic tradition. Bellow we present briefly our approach to the encoding of the semantic relations in BulNet.

2.1 SYNONYMY

SYNONYMY is a semantic relation of equivalence (reflexive, symmetric, and transitive) between literals belonging to one and the same part of speech - *if A is synonymous to B, B is synonymous to A*. The synonyms (one or more) form the synonym set - so called synset. In Princeton WordNet the substitution criteria for SYNONYMY is mainly adopted: "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" [5]. Thus the relation implies that one synonym may substitute another (synonym) in a context and vice versa. The consequences from such an approach are at least two - not only the exact SYNONYMY is included in the data base (a context is not every context). Second, it is easy to find contexts in which words are interchangeable, but still denoting different concepts (for example hypernyms and hyponyms), and there are many words which have similar meanings and by definition they are synonyms but are hardly interchangeable in any context due to different reasons - syntactic, stylistic, etc. (for example an obsolete and a common word). The standard implicative tests of the following type could be used for SYNONYMY identification [2] [9]:

Word1 in a Context C entails and is entailed by Word2 in a Context C.

Since neither defining the SYNONYMY in terms of substitution, nor - in terms of semantic similarity are reliable enough in practice (developing BulNet), the term SYNONYMY is generally used to refer to the relation between literals that: a) belong to the same POS - obligatory; b) have the same semantic meaning - obligatory; c) could be interchangeable in a context - optional; d) could belong to the same semantic field - optional; e) could have equivalent obligatory syntactic environments - optional.

2.2 HYPERNYMY and HYPONYMY

HYPERNYMY and HYPONYMY are inverse, asymmetric, and transitive relations between synsets, which correspond to the notion of class-inclusion: *if W1 is a kind of W2, then W2 is hypernym of W1 and W1 is a hyponym of W2.* The relation implies that the hypernym may substitute for the hyponym in a context but not the other way round. HYPONYMY / HYPERNYMY is a transitive relation: e.g. being a kind of *cvete (flower), roza (rose)* has inherited not only all semantic features of *cvete (flower)*, but also those of its superordinates: *rastenie (plant), zhiv organizam (organism)*, etc. HYPERNYMY and HYPONYMY relations are between synsets belonging to one and the same part of speech. The cross part-of-speech HYPERNYMY is subject to different relation, named XPOS HYPERNYMY [9]. The general implicative test sentence for HYPONYMY is unidirectional:

Word1 in a Context C entails but is not entailed by Word2 in a Context C.

In BulNet we allow the existence of multi-parent relations, i.e. one synset may have several hypernyms, e.g., *aktrisa (female actor)* is subordinated to both *actor* and *woman.* Though it is rather difficult to deal with multiple inheritances within databases, we believe that it helps us to reflect the real structure of semantic relations within the language. Multiple hypernyms have occasionally been encoded in WordNets. In the English database WordNet 1.5 only 582 synsets are in relation with two hypernyms. A *knife* can be either a *weapon* or a *piece of cutlery.* Thus the word *knife* should receive different definitions being weapon or piece of cutlery and should belong to two different synsets. A *spoon* could be both a *container* and a *piece of cutlery.* Thus *spoon* should have two hypernyms, and consequently should belong to two different paths of the HYPERNYMY tree, but the two paths should have a common hypernym somewhere in the top structure. The hyponym inherits a semantic meaning both from its first hypernym and its second hypernym (conjunction is applied) and thus from the higher hypernyms, too.

The results of such approach should be avoiding of some artificial hierarchies between words. However, the correspondence with the English WordNet structure would remain.

In BulNet the term HYPERNYMY is generally used to refer to the relation between synsets that: a) belong to the same POS – obligatory; b) have some common semantic component(s) – obligatory; c) are interchangeable in one direction - obligatory. The HYPERNYMY relation may be verified

with the help of some other relations: NEAR ANTONYMY, MERONYMY, etc. as it is shown below.

2.3 NEAR ANTONYMY

ANTONYMY is a symmetric and intransitive relation of opposition (it is established between two members, belonging to one and the same part of speech) — *if A is antonymous to B, B is antonymous to A.* It is disputable whether ANTONYMY stands between either word forms or synsets (word meanings). The solution adopted by Miller's WordNet is that ANTONYMY is considered to be a relation between word forms, but not between word meanings. In the cases that ANTONYMY holds for all members of the synset, a separate NEAR ANTONYMY relation is used. Another solution of this problem is reordering of synset members: usually one synset representative (the dominant literal) is related by ANTONYMY directly; all other members of synsets are opposed through this pair, i.e. indirectly.

For verification of NEAR ANTONYMY we apply bidirectional implicative test sentences with negation:

Word1 in a Context C entails not-Word2 in a Context C.

Word2 in a Context C entails not-Word1 in a Context C.

In BulNet the NEAR ANTONYMY relation is defined between synsets that: a) belong to the same POS - obligatory; b) have some common semantic component(s) - obligatory; c) are opposed by some essential semantic component(s) - obligatory; d) belong to the same semantic field - optional; e) are interchangeable in a context with negation - optional. Other criteria for verification can be formulated with the help of other semantic relations: a) antonyms have to share the same hypernym (not compulsorily the immediate one) - obligatory; b) hyponyms of two antonyms (nouns) should also be antonyms pair by pair *(woman - man; female actor - actor)* - obligatory.

2.4 MERONYMY and HOLONYMY

MERONYMY and HOLONYMY are asymmetric relations which link synsets denoting wholes with those denoting their parts: *if W1 has a W2, and W2 is a part, portion, member of W1, then W1 is a meronym of W2 and W2 is a holonym af W1.* It is noticed that MERONYMY may not be always reversible to HOLONYMY: e.g., whereas a forest is not a forest unless it consists of trees, a tree does not necessarily grow in a forest (it may be in a street or in a desert) [2]. Unlike hyponymy, transitivity of MERONYMY

is limited - some cases of MERONYMY are not transitive. This is due to the fact that within the frame of MERONYMY we can distinguish at least five different types of relations. Though transitive meronyms can sometimes be arranged in hierarchical structures, usually they are incorporated into net-like structures rather than trees (e.g. *eye* may be a part of *face*, as well as a part of *visual system*; *face* may be a part both of *human*, as well as part of *head*; *head* may be part of *body* and *animal* - thus we can establish multiple MERONYMY relations.

The following types of MERONYMY can be distinguished:
- PART OF: *klon - darvo (branch - tree)*;
- MEMBER OF: *darvo - gora (tree - forest)*;
- PORTION OF: *parche torta - torta (a piece of cake - cake)*;
- SUBSTANCE OF: *aluminij - samolet (aluminum - airplane)*;
- LOCATION OF: *oazis - pustinja (oasis - desert).*

The first three types of MERONYMY are included in BulNet.

2.4.1 PART MERONYMY: MERO PART and HOLO PART

The has holo/mero part relation typically relates components to their wholes. Namely, something which is either topologically or temporally included in a larger entity and which as well bears some kind of autonomy (non-arbitrary boundaries) and a definite function with respect to the whole [9]. In BulNet we restrict the has holo/mero part relation to the components that are topologically included one in the other with physical attachment: *book* is a part of *library*, *library* is a part of *building*, **book* is not a part of *building*, only *library - building* relation is encoded as PART MERONYMY. That is why in our point of view has holo/mero part are inverse asymmetric transitive relations between noun synsets, which link objects denoting wholes with those denoting their parts: *if W1 has a W2, and W2 is a part of W1 then W1 is a mero part of W2 and W2 is a holo part of W1.* The condition states that there must be multiple components (which can be of the same type) and that both the holonym and the meronym should be concrete objects. Restrictions on topological inclusion and physical attachment predetermine reversion and transitivity (if there are more than two hierarchical components, which is a rare case) of the relation. If we consider the relation *branch – tree – forest*, between *branch* and *tree* a PART MERONYMY is established, while between *tree* and *forest* exists MEMBER MERONYMY. The holo/mero parts build structures that rarely consist of more than two members, if they do, the holo/mero PART MERONYMY is transitive: *room*

– *apartment – block of flats*, the *room* is a component of a *block* and *block* consists of *rooms*.

For verification of PART MERONYMY we apply unidirectional implicative test sentences:

Word1 in a Context C is a part of Word2 in a Context C.

**Word2 in a Context C is a part of Word1 in a Context C.*

In BulNet the PART MERONYMY is between synsets that: a) are concrete nouns – obligatory; b) denote wholes and their parts – obligatory; c) belong to the same semantic field - optional. The criteria for verification based on the correspondence with other relations are: a) hyponym should have the same mero parts (for concrete nouns) as the hypernym *(man - head, arm, ; woman - head, arm, ..)* – obligatory; b) antonyms should have equivalent holo parts: *(woman - arm, head; man - arm, head* – obligatory.

2.4.2 MEMBER MERONYMY: MERO MEMBER and HOLO MEMBER

The has holo/mero member relation is between sets and their members: it is inverse asymmetric transitive relations between noun-synsets: *if W1 has a W2, and W2 is a member of W1 than W1 is a mero member of W2 and W2 is a holo member af W1.* The condition states that holo member word is a single object denoting noun and mero member word is a multiform noun (either a group-noun, a collective-noun or as a lexicalized plural denoting multiple objects). The MEMBER MERONYMY builds structures that rarely consist of more than two synsets, if it does the holo/mero PART MERONYMY is transitive and the higher binary relation is between two collective nouns: *football player - football team - football league*, *football player* is a member of a *football team*, *football team* is a member of *football league*, as well as a *football player* is a member of *football league*.

For verification of MEMBER MERONYMY we apply unidirectional implicative test sentences:

Word1 in a Context C is a member of Word2 in a Context C.

**Word2 in a Context C is a member of Word1 in a Context C.*

In BulNet the relation MEMBER MERONYMY is used to describe the relation between synsets that: a) are concrete and collective nouns or only collective nouns – obligatory; b) denote sets (collective nouns) and their members (concrete nouns or collective nouns) – obligatory; c) belong to the same semantic field - optional. Other criteria for verification based on

the correspondence with other relations are: a) collective nouns that are holo/mero members should share the same hypernym, not compulsorily the immediate one (*football team* is an *organization*, as well as *football league*) - obligatory.

2.4.3 PORTION MERONYMY: MERO PORTION and HOLO PORTION

The has holo/mero portion relation is between wholes and their portions: it is an inverse asymmetric transitive relation between noun-synsets: *if W1 has a W2, and W2 is a portion of W1 than W1 is a mero portion of W2 and W2 is a holo portion of W1.* The condition states that the whole (noun denoting substance, material) always presupposes the portion and usually portions (as concepts) do not receive a separate lexical item but are realized by sense extension (for instance, there is no lexical item in Bulgarian equivalent to *"portion of cake").* The PORTION MERONYMY creates structures that rarely consist of more than two synsets, if it does the holo/mero PORTION MERONYMY is transitive and lowest binary relation is between nouns denoting portions: *crumb of bread - slice of bread - loaf of bread*; *crumb of bread* is a portion of a *slice of bread*, *slice of bread* is a portion of *loaf of bread*, as well as *crumb of bread* is a portion of *loaf of bread.*

For verification of PORTION MERONYMY we apply unidirectional implicative test sentences of the following type:

Word1 in a Context C is a portion of Word2 in a Context C.

**Word2 in a Context C is a portion of Word1 in a Context C.*

In BulNet the term PORTION MERONYMY refers to the relation between synsets that: a) are nouns - obligatory; b) denote wholes (expressing substance, material) and their portions - obligatory; c) belong to the same semantic field - optional.

Other criteria for verification based on the correspondence with other relations are: a) nouns that are holo/mero portions should share the same hypernym, not compulsorily the immediate one (*coffee - substance; caffeine - substance*) - obligatory.

2.5 HAS SUBEVENT and IS SUBEVENT OF

ENTAILMENT is one of the important relations specific mostly for verbs and their derivatives [3]. Two basic kinds of lexical entailment were distinguished: one involves 'temporal inclusion' (the two situations referred to

by the verbs in the relation partially or totally overlap); the other involves 'temporal exclusion' (the two situations are variously temporally disjoint):

a) + Temporal Inclusion: a.1) co-extensiveness (*to limp - to walk*)' a.2) proper inclusion (*to snore - to sleep*);

b.) - Temporal Exclusion: b.1) backward presupposition (*to succeed - to try*), b.2) causation (*to give - to have*).

In BulNet data related to the WordNet2.0 ENTAILMENT relations are encoded as follows: (a1) is referred to as HYPONYMY, (a2) is referred to as SUBEVENT, (b1) and (b2) are referred to as CAUSATION.

HAS SUBEVENT and IS SUBEVENT OF are inverse, asymmetric and (intransitive) relations between verb synsets, one event takes place during or as a part of the another, and whenever the first event takes place, the second one also takes place: *if W1 and W2 take place simultaneously, and W2 takes place as a part of W1, than W2 is SUBEVENT of W1.*

For verifying SUBEVENT relations we use test sentences of the following type:

Word1 in a Context C has subevent Word2 in a Context C.

**Word2 in a Context C has subevent Word1 in a Context C.*

Within BulNet we apply the term SUBEVENT to the entailment relations between synsets that: a) belong to the same POS (verbs) – obligatory; b) refer to two temporally conjoint situations – obligatory; (c) refer to the complex activity and its simple part - obligatory; d) referred situations have the same agent – obligatory; e) referred situations have equivalent syntactic environment - optional.

2.6 CAUSE and IS CAUSED BY

CAUSE and IS CAUSED BY are inverse, asymmetric, and intransitive relations between verb synsets, one of the synsets refers to an event causing another event, process or state referred to by the second synset: *if W1 causes W2, then W2 is caused by W1.*

The causal relation is used for verb pairs such as *show/see, fell/fall, give/have.* For verifying cause relations we use test sentences of the following type:

Word1 in a Context C may cause Word2.

**Word2 in a Context C may cause Word1.*

The term CAUSATION is used to indicate the entailment relation between synsets that: a) belong to the same POS (verbs) – obligatory; b) refer to temporally disjoint situations – obligatory; c) one of the synsets refers to

an event causing another event, process or state referred to by the second synset – obligatory; d) events referred have different agents - optional; e) referred situations have equivalent syntactic environment - optional.

2.7 SIMILAR TO

SIMILAR TO is a symmetric relation between synsets belonging to the POS adjective: *if Word1 is similar to Word2, then Word2 is similar to Word1.* The relation can be described as relation of semantic similarity [5] between a focal synset which has a given referential meaning and synsets that have close referential meaning and/or different stylistic, connotative, etc. features or express different degree of a certain state or quality (such as *beautiful and pretty*). The adjective synsets involved in the SIMILAR TO relation form clusters organized around the focal synset. The concept of the focal synset also accounts for antonymy [5]. A lot of adjectives cannot be said to have antonyms, although they have a conceptual opposite. Through SIMILAR TO such adjectives are indirectly linked to their opposites.

We can distinguish the SIMILAR TO relation via: a) the two synsets should have the same values for the POS tag (adjective) – obligatory; b) the adjective synsets should be related to a focal synset which mediates their semantic relations with other synsets – obligatory.

2.8 VERB GROUP

VERB GROUP is a symmetric relation between semantically related verb synsets; *if Word1 belongs to the verb group of Word2, then Word2 belongs to the verb group of Word1.* The relation presupposes lexical relation as well: the two synsets interrelated through VERB GROUP have common semantic component (lexicalized by a common literal) which is present in the meaning of all the synsets that enter in the VERB GROUP relation. The semantic difference may be accompanied by change in the semantic structure of the verbs which can account for the difference in their subcategorization. As with SIMILAR TO there is a focal synset which provides the common semantic basis for the cluster of synsets.

We can distinguish the VERB GROUP relation via: a) the two synsets should have the same values for the POS tag (verb) – obligatory; b) the verb synsets should be related to a focal synset with common semantic meaning - obligatory; c) verbs have to have equivalent syntactic environment - optional.

2.9 AlSO SEE

AlSO SEE is a symmetric relation between synsets with the same POS value, provided it is verb or adjective. The relation holds between synsets that are close in meaning, but unlike the SIMILAR TO relation where a head synset has the function of mediating between words that are very close in meaning, but are not systematically related to other synsets through semantic relations, with ALSO SEE each synset entering this type of relation has its own systematic relations with other synsets: its own SIMILAR TO, NEAR ANTONYM, etc. relations.

ALSO SEE differs from the VERB GROUP relation of verb synsets as well. While the verb group relation may be said to encode the phenomenon of verb polysemy, ALSO SEE accounts for a further differentiation in meaning which is also observable on the lexical level which results in close, but not equivalent literals in the synsets.

3 WordNet Logic

For the exploring, using and analyzing of the huge WordNet data base, a powerful querying system is required. A specific system for a predefined set of queries and a Prolog encoding of WordNet were developed for this purposes in Princeton `http://www.cogsci.princeton.edu/wn/`. Another ontological description of the WordNet noun and verb synsets is created at Suggested Upper Merged Ontology `http://ontology.teknowledge.com`.

In the next section we introduce the formal definition of the WordNet structure and define the syntax and semantics of our formal language [8]. Afterwards we proceed with the WordNet Logic in Section 3.2. The logic completeness and decidability results are presented in Section 3.3.

3.1 Syntax and Semantics

3.1.1 WordNet structure

We call a WordNet structure the tuple:
$\mathcal{F} = \langle W, \equiv, \mathbf{Hyp}, \mathbf{Ili}, \mathbf{Lit}, \mathbf{Lang}, \mathbf{Base}, \mathbf{Glos}, \mathbf{POS} \rangle$, where:

- W is the set of word senses
- $\equiv \subset W^2$ is the synonymy relation
- $\mathbf{Hyp} \subset W^2$ is the direct hyperonymy relation

- **Ili**$\subset W^2$ is the Inter-Lingual Index relation
- **Lang**$\subset W^2$ is the language relation
- **Lit**$\subset W^2$ is the literal relation
- **POS** $\subset 2^W$ is the part of speech splitting of W
- **Glos** $\subset W$ is the "defined gloss" property
- **Base** $\subset W$ is the "base concept" property,

if the following conditions hold:

- $W \neq \emptyset$
- $\equiv$, **Ili**, **Lit**, **Lang** are equivalence relations on W and $\equiv$ is the intersection of **Ili** and **Lang** i.e. $\equiv$=**Ili** $\cap$ **Lang**
- **Hyp**$\subset W^2$ is irreflexive i.e. $\forall x \in W(\neg x\ \mathbf{Hyp}\ x)$
- **Hyp** is coherent in respect to $\equiv$ i.e.

$$\forall x, y, z \in W(((x \equiv y)\ \&\ (x\ \mathbf{Hyp}\ z)) \rightarrow (y\ \mathbf{Hyp}\ z))$$
$$\forall x, y, z \in W(((x \equiv y)\ \&\ (z\ \mathbf{Hyp}\ x)) \rightarrow (z\ \mathbf{Hyp}\ y))$$

- **POS** $= \{P_N, P_V, P_{Adj}, P_{Adv}\}$ and $\bigcup$ **POS** $= W$ and the elements of **POS** are disjoint and coherent in respect to **Ili** and **Hyp** i.e.

$$\forall \mathbf{P} \in \mathbf{POS}\, \forall x, y \in W(((x\ \mathbf{Ili}\ y)\ \&\ (x \in \mathbf{P})) \rightarrow (y \in \mathbf{P}))$$

$$\forall \mathbf{P} \in \mathbf{POS}\, \forall x, y \in W((x\ \mathbf{Hyp}\ y) \rightarrow ((x \in \mathbf{P}) \leftrightarrow (y \in \mathbf{P})))$$

- **Base** and **Glos** are coherent in respect to $\equiv$ i.e.

$$\forall x, y \in W(((x \equiv y)\ \&\ (x \in \mathbf{Base})) \rightarrow (y \in \mathbf{Base}))$$
$$\forall x, y \in W(((x \equiv y)\ \&\ (x \in \mathbf{Glos})) \rightarrow (y \in \mathbf{Glos})).$$

Let us now define the language over the WordNet structure.

3.1.2 Syntax of the $\mathcal{WN}$ Language

The set of relation symbols is $Rel = \{\equiv, Hyp, Hyp^+, \widetilde{Hyp}, \widetilde{Hyp}^+, Ili, Lit, Lang, \circ\}$.

The set of variable types is: $VT = \{0, \equiv, Lit, Lang, Ili, WS\}$

The set of variables is: $Prop = \{p^\tau \mid \tau \in VT, p \text{ ranges over an infinite symbol set}\}$

We have a finite number of constant symbols for the types $Lit, Lang, Ili$ labeled in correspondence with their meaning. For example for type Lit we shall use constant names *cat*, *chair*, *Wörterbuch* etc. For type $Lang$ we shall use *English*, *Bulgarian* etc. For type Ili we shall use natural numbers like 00002031.

We have property constants for the part of speech classes – $P_N, P_V, P_{Adj}, P_{Adv}$, for the base concept – $Base$, and for the defined gloss property – $Glos$.

The *Elementary propositions* over $\mathcal{WN}$ are the variables and constants of all types defined above.

We define inductively the formulae of $\mathcal{WN}$:

- The Elementary propositions are formulae
- if φ and ψ are formulae, then:

  $$(\varphi \,\&\, \psi), (\varphi \vee \psi), (\varphi \rightarrow \psi), (\varphi \leftrightarrow \psi), (\neg\varphi)$$

  are formulae
- if φ is a formula and $R \in Rel$ is a relation symbol, then:

  $$([R]\varphi), (\langle R\rangle\varphi)$$

  are formulae.

3.1.3 Semantics of the $\mathcal{WN}$ Language

A WordNet structure is called appropriate for a given $\mathcal{WN}$ language if there exists a function which maps the set of property constants to the distinguished subsets of W corresponding to their meaning. I.e. the function has to map $P_N, P_V, P_{Adj}, P_{Adv}, Base, Glos$ to the corresponding subsets of the WordNet structure.

Valuation V is a mapping of the variables and non-property constants to subsets of W, which preserves the types. More precisely $V(p^0)$ is an arbitrary subset of W; $V(p^{\equiv})$ is an equivalence class of $\equiv$; $V(p^{Lit})$ is an equivalence class of **Lit**; $V(p^{Lang})$ is an equivalence class of **Lang**; $V(p^{Ili})$

is an equivalence class of **Ili**; $V(p^{WS})$ is a singleton of W. The valuation of the constants has to follow the same type constraints.

A $\mathcal{WN}$ model over an appropriate WordNet structure $\mathcal{F}$ is a couple $\langle \mathcal{F}, V \rangle$, where V is a valuation.

Here we shall define the interpretation of the $\mathcal{WN}$ language over a WordNet structure.

In the structure $\mathcal{F}$ with $\mathbf{Hyp}^+, \widetilde{\mathbf{Hyp}}, \widetilde{\mathbf{Hyp}}^+$ we denote the transitive closure, the inverse and the transitive closure of the inverse relation of **Hyp** respectively.

We define the truth of formula φ of a $\mathcal{WN}$ language in the point $x \in W$ over the model $\langle \mathcal{F}, V \rangle$ by induction on the formula construction:

1. $x \Vdash PC$ iff x is an element of the corresponding property subset in $\mathcal{F}$, where $PC \in \{P_N, P_V, P_{Adj}, P_{Adv}, Base, Glos\}$
2. $x \Vdash p$ iff $x \in V(p)$, where p is a variable or non-property constant
3. $x \Vdash \neg\varphi$ iff $x \nVdash \varphi$
4. $x \Vdash (\varphi \,\&\, \psi)$ iff $x \Vdash \varphi$ and $x \Vdash \psi$
5. $x \Vdash (\varphi \vee \psi)$ iff $x \Vdash \varphi$ or $x \Vdash \psi$
6. $x \Vdash (\varphi \rightarrow \psi)$ iff $x \Vdash \varphi \Rightarrow x \Vdash \psi$
7. $x \Vdash (\varphi \leftrightarrow \psi)$ iff $x \Vdash \varphi \Leftrightarrow x \Vdash \psi$
8. $x \Vdash ([R]\varphi)$ iff $\forall y \in W(x\mathbf{R}y \Rightarrow y \Vdash \varphi)$, where $R \in Rel$
9. $x \Vdash (\langle R \rangle \varphi)$ iff $\exists y \in W(x\mathbf{R}y$ and $y \Vdash \varphi)$, where $R \in Rel$

A formula φ of a $\mathcal{WN}$ language is true over a model $\langle \mathcal{F}, V \rangle$ if it is true for every point $x \in W$.

3.2 WordNet Logic

3.2.1 Axiomatic System

Axioms for normal modal logic

- All tautologies of the classic propositional logic
- $[R](\varphi \rightarrow \psi) \rightarrow ([R]\varphi \rightarrow [R]\psi)$, where $R \in Rel$

Axioms for $\circ$

- $\varphi \rightarrow \langle\circ\rangle\varphi$
- $\langle\circ\rangle\varphi \rightarrow \langle\circ\rangle\langle\circ\rangle\varphi$
- $\varphi \rightarrow [\circ]\langle\circ\rangle\varphi$
- $\langle R\rangle\varphi \rightarrow \langle\circ\rangle\varphi$, where $R \in Rel$

Axioms for $\equiv, Ili, Lit, Lang$

- $\varphi \rightarrow \langle R\rangle\varphi$, where $R \in \{\equiv, Ili, Lit, Lang\}$
- $\langle R\rangle\varphi \rightarrow \langle R\rangle\langle R\rangle\varphi$, where $R \in \{\equiv, Ili, Lit, Lang\}$
- $\varphi \rightarrow [R]\langle R\rangle\varphi$, where $R \in \{\equiv, Ili, Lit, Lang\}$

Axioms for coherence

- $\langle\equiv\rangle p^{WS} \leftrightarrow \langle Ili\rangle p^{WS}$ & $\langle Lang\rangle p^{WS}$
- $p^{WS} \rightarrow [Hyp]\neg p^{WS}$
- $\langle\equiv\rangle p^{WS}$ & $\langle Hyp\rangle q^{WS} \rightarrow \langle\equiv\rangle(p^{WS}$ & $\langle Hyp\rangle q^{WS})$
- $P_N \vee P_V \vee P_{Adj} \vee P_{Adv}$
- $P_i \rightarrow \neg P_j$, where $i, j \in \{V, N, Adj, Adv\}, i \neq j$
- P_i & $\langle Ili\rangle p^{WS} \rightarrow \langle Ili\rangle(P_i$ & $p^{WS})$, where $i, j \in \{V, N, Adj, Adv\}$
- P_i & $\langle Hyp\rangle p^{WS} \rightarrow \langle Hyp\rangle(P_i$ & $p^{WS})$, where $i, j \in \{V, N, Adj, Adv\}$
- $Base$ & $\langle\equiv\rangle p^{WS} \rightarrow \langle\equiv\rangle(Base$ & $p^{WS})$
- $Glos$ & $\langle\equiv\rangle p^{WS} \rightarrow \langle\equiv\rangle(Glos$ & $p^{WS})$

Axioms for Hyponymy and transitive closure

- p^{WS} & $\langle Hyp\rangle q^{WS} \rightarrow \langle Hyp\rangle(q^{WS} \rightarrow \langle\widetilde{Hyp}\rangle p^{WS})$
- p^{WS} & $\langle\widetilde{Hyp}\rangle q^{WS} \rightarrow \langle\widetilde{Hyp}\rangle(q^{WS} \rightarrow \langle Hyp\rangle p^{WS})$
- p^{WS} & $\langle Hyp^{+}\rangle q^{WS} \rightarrow \langle Hyp^{+}\rangle(q^{WS} \rightarrow \langle\widetilde{Hyp}^{+}\rangle p^{WS})$

- p^{WS} & $\langle \widetilde{Hyp}^{+} \rangle q^{WS} \rightarrow \langle \widetilde{Hyp}^{+} \rangle (q^{WS} \rightarrow \langle Hyp^{+} \rangle p^{WS})$
- $\underbrace{\langle Hyp \rangle \langle Hyp \rangle \cdots \langle Hyp \rangle}_{i+1} p^{WS} \rightarrow \langle Hyp^{+} \rangle p^{WS}$, for all $i \in$

Axioms for variable types

- $\langle \circ \rangle p^{WS}$
- $\langle \circ \rangle (p^{WS}$ & $\varphi) \rightarrow [\circ](p^{WS} \rightarrow \varphi)$ for all formulae φ
- $p^{R} \rightarrow [R]p^{R}$ for $R \in \{\equiv, Ili, Lang, Lit\}$
- p^{R} & $\langle \circ \rangle (q^{WS}$ & $p^{R}) \rightarrow \langle R \rangle q^{WS}$ for $R \in \{\equiv, Ili, Lang, Lit\}$

Inference rules

- (MP)

$$\frac{\varphi, \varphi \rightarrow \psi}{\psi}$$

- (Nec)

$$\frac{\varphi}{[\circ]\varphi}$$

We need to define the notion of *Normal Admissible Form* in order to present the next two rules. We define *Normal Admissible Form* by induction:

1. # is a Normal Admissible Form
2. if φ is a formula and A is a Normal Admissible Form, then $(\varphi \rightarrow A)$ is a Normal Admissible Form
3. if A is a Normal Admissible Form and $R \in Rel$ is a modality, then $([R]A)$ is a Normal Admissible Form

Let us note that every Normal Admissible Form contains exactly one occurrence of # and the result of the substitution of # in a Normal Admissible Form A with a formula φ – $A(\varphi)$ – is a formula.

- (Cov)

does not occur in AA$(\varphi$ & $\neg\varphi)$(Ind$^{\infty}$)

$$\frac{A(\underbrace{[Hyp][Hyp] \cdots [Hyp]}_{i+1} \varphi) \text{ for all } i \in \text{, where } A \text{ is a Normal Admissible Form}}{A([Hyp^{+}]\varphi)}$$

The set of theorems of the $\mathcal{WN}$ logic is the smallest set of $\mathcal{WN}$ formulae, which contains the axioms and is closed under the inference rules.

3.3 WordNet Logic Soundness, Completeness, and Decidability Theorems

[Soundness] Every theorem is true in every $\mathcal{WN}$ model.

[Completeness] If φ is a formula and φ is true in every $\mathcal{WN}$ model, then φ is a theorem.

There exists an algorithm, which for a given $\mathcal{WN}$ formula, finishes after a finite number of steps and if the formula is satisfiable returns a model and a point where the formula is satisfiable. Otherwise the algorithm returns "false".

Satisfiability problem for $\mathcal{WN}$ formulae is decidable.

It is decidable whether a given formula is a theorem.

The proofs of the above results can be completed using technique similar to the one presented in [6, 1, 7].

4 Completeness of the Bulgarian WordNet

The main positive characteristics of the BulNet are its completeness and consistency. In our work we successively used the predefined WordNet Logic queries for validation of our data.

Under completeness we understand the presence of all members from the Base Concepts chosen up to now within the framework of the BalkaNet project. These are Base Concepts subset 1 (1 218 synsets), Base Concepts subset 2 (3 471 synsets) and Base Concepts subset 3 (3 789 synsets), that are encoded in the electronic databases of the languages involved in the BalkaNet project: Bulgarian, Czech, Greek, Romanian, Serbian and Turkish. We measure the completeness of the BulNet not only by the number of the common synsets in all languages but also according to several additional criteria: lack of any "dangling relations" in the data base - that is both members of the defined relation have to be present in the WordNet; lack of any "gaps" - if a certain synset is included in the Bulgarian WordNet, then all of its hypernyms should be present up to the top of the tree; lack of any "free" nodes - a synset included in a WordNet should be in a relation at least with one different synset. Each synset must contain at least one literal, as well as at least one language-internal relation must be defined for each synset.

Finally, we consider the WordNet complete if the following tags have received a value: the synset ID tag which makes the relation to the corresponding synset in English WordNet2.0 explicit, the synset POS tag ensuring that each synset is specified for the part of speech it belongs to, the synset DEF - an appropriate interpretation definition must be entered for each synset, the SENSE tag - each literal has to receive unique sense number that distinguishes it from the homographic literals with different meaning, the synset BCS tag - each synset has to be defined as to whether or not it belongs to a particular Base Concept subset. On the other hand, there are some XML tags such as USAGE, SNOTE, LNOTE, STAMP which are not obligatory, so they may not possess a value and are removed automatically if empty. The completeness of the current state of the BulNet can be exemplified with the following Table 1:

NUMBER OF SYNSETS	18810
NUMBER OF LITERALS	34266
BASE CONCEPTS SUBSET 1	1218
BASE CONCEPTS SUBSET 2	3471
BASE CONCEPTS SUBSET 3	3789
EMPTY TAGS	0
SYNSETS WITHOUT ID TAG VALUE	0
SYNSETS WITHOUT POS TAG VALUE	0
SYNSETS WITHOUT BCS TAG VALUE	0
SYNSETS WITHOUT DEFINITION	0
SYNSETS WITHOUT LITERALS	0
SYNSETS WITHOUT ILR	0
"FREE" SYNSETS	0
"DANGLING"	0
"GAPS"	0
LITERALS WITHOUT SENSE TAG VALUE	0

-

Table 1: Statistics for the completeness of Bulgarian WordNet

5 Consistency of the Bulgarian WordNet

The second important characteristic of the BulNet is its consistency. Our approach to the validation of the Bulgarian WordNet comprises three separate steps of different degrees of complexity and significance that present different possibilities for automatic data correction: checking the consistency of the syntax of the XML files in which the data are organized, checking for contradictions in the interpretation meanings of the synsets, and checking the consistency in the encoding of the semantic relations.

The lowest level, which is also the easiest for processing and correction, is XML files syntax. We apply an automatic checking as well as automatic data correction in the following cases:

– The literals in a given synset have to be unique, thus the duplicated literals are eliminated automatically while keeping at least one of them.

– Currently, sense numbers are random in the Bulgarian WordNet; they do not correspond to the arrangement of the meanings of polysemy words in an explanatory dictionary or to the frequency of usage of a certain meaning. We provide the required checking (and automatic reordering) of the SENSE tags, to ensure that each tag possesses a value, that contains only numbers, as well as that the sense numbers are successive and are not duplicated.

In other cases where automatic correction of consistency is possible, manual confirmation of replacements is absolutely necessary:

– The identification tags are checked to verify whether they conform to the accepted standard - a certain number of digits and part of speech denotation, and (if they do not match) the closest match in the English WordNet 2.0 is suggested. Of course, the automatic replacement without checking should be avoided and the decisions for the correct ID connection are to be taken by a lexicographer.

– To the Part of Speech tags and Base Concepts tags, whose values differ from the corresponding English ones, respective English tag values are automatically assigned. Manual confirmation of the replacement is again compulsory in these cases, because there are examples (rare as they might be) where the English and Bulgarian translation equivalents belong to different parts of speech.

The third option is an automatic validation and manual correction of missing or incorrect parts of the XML file:

– The text sections in the Bulgarian XML file should include only Cyrillic characters. Of course, mistakes are possible, where a Cyrillic "а" is replaced by a Latin "a", or where parts that have not been translated from English

have been kept. These errors must be checked and if necessary corrected. It is clear, however, that the Latin characters must be kept in certain cases, such as chemical elements denotations, Latin names of plants, animals, etc. A continuation of this task is spelling checking of Bulgarian text parts.

– For ID tags a check is performed to find out whether there are empty tags or duplicated ID numbers. Correspondingly, the decision whether the ID tag is correctly connected or should be removed (if duplicated) is to be taken by a human expert.

– Another important verification is the checking for duplicated internal language relations between two synsets in a language. A synset can be connected with an arbitrary number of language internal relations if and only if each relation links the synset to different synsets in the data base. The duplication of relations is detected and a lexicographer takes a decision which relation is correct and which is to be removed.

As a result of the application of the specified methodology for checking and correction of the Bulgarian WordNet, the current status of the XML syntax is the following (Table 2):

DUPLICATED LITERALS IN A SYNSET	0
DUPLICATED SENSE NUMBERS	0
INCONSEQUENT SENCE NUMBERS	0
MISSING SENSE NUMBERS	0
DEFECTED ID TAGS VALUES	0
DEFECTED POS TAGS VALUES	0
DEFECTED BCS TAGS VALUES	0
SPELLING ERRORS	0
WORDS IN LATIN CHARACTERS	961
EMPTY ID'S	0
DUPLICATED SYNSETS	0
DUPLICATED RELATIONS	0

Table 2: The consistency of the Bulgarian XML file

The most difficult and important task is the verification of the consistency of the data itself - the semantic relations and the interpretation meanings of the synsets. When validating relations already defined for a given synset the following tests are used:

– All Bulgarian synsets whose hypernym differs from the English ones and synsets without a hypernym are checked again. This check is broadened to cover all relations. Every difference in relations between EWN2.0 and the Bulgarian WordNet is either language specific and linguistically substantiated or is due to the fact that one of the synsets is not yet presented in the WordNet.

– There must be no hypernym cycles, as well as any relation loops inside WordNet. The cycle is defined easily in such (artificial) examples like following:

"Rose" has hypernym "flower".

"Flower has hypernym rose". (one step cycle)

It is not clear how to define the errors in cases of multiple hypernymy (or any transitive) relation - e.g., *"eye"* may be a part both of *"face"* as well as a part of *"visual system"*, *"face"* may be a part both of *"human"* as well as part of *"head"*, *"head"* may be part of *"body"* and *"animal"*. A similar example concerning hypernymy is:

"Oxygen" has hypernym "gas".

"Gas" has hypernym "fluid".

"Fluid" has hypernym "substance".

"Oxygen" has hypernym "chemical element".

"Chemical element" has hypernym "substance".

In some cases there are wrongly connected nodes, but some cases may be instances of different subrelations. For example, the distinction between the following types of hyponymy is not included for the time being in the Bulgarian WordNet: *"kingdom"* is a kind of *"state"*, while *"Bulgaria"* is an instance of *"state"*; *"actor"* is a role of *"person"*, while *"man"* is a type of *"person"*. If we allow such subrelations, we could avoid multiple transitive relations for a synset and thus we could successfully apply the consistency validation.

When checking for glosses' consistency the following tests may be used:

– It is automatically checked whether there are literals in the Bulgarian WordNet that coincide with their glosses. In such cases the glosses are redefined.

– Another check is whether the glosses of different synsets are identical and if they are – the interpretation definitions are compared and differentiated in an appropriate manner.

– When building the Bulgarian WordNet, we have come across the problem of English synsets that denote concepts existing in the Bulgarian language consciousness but are not lexicalized in Bulgarian. In such cases we

have adopted the strategy of keeping the node in the Bulgarian WordNet and marking it with the phrase "no lexicalization". At the moment we have 99 language specific concepts defining relative relations such as *"baldaza"* (the sister of one's wife) and some adjectives. The next table illustrates the level of the consistency in the Bulgarian WordNet (differences in the relations does not involve inconsistency).

DIFFERENCE IN ID's	99
EQUIVALENT GLOSSES	0
GLOSSES EQUAL WITH LITERALS	0
DIFFERENCE IN RELATIONS hypernym	19
DIFFERENCE IN RELATIONS be in state	16
DIFFERENCE IN RELATIONS also see	365
DIFFERENCE IN RELATIONS similar to	454
DIFFERENCE IN RELATIONS holo part	68
DIFFERENCE IN RELATIONS holo member	10
DIFFERENCE IN RELATIONS subevent	0
DIFFERENCE IN RELATIONS causes	0
DIFFERENCE IN RELATIONS derived	77
DIFFERENCE IN RELATIONS particle	0
DIFFERENCE IN RELATIONS verb group	21
DIFFERENCE IN RELATIONS near antonym	9
DIFFERENCE IN RELATIONS holo portion	9
ANY LOOPS	0

Table 3: The consistency of the encoded relations and definitions

6 Conclusions

All relations included in the BulNet structure are carefully examined and (if necessary) predefined according Bulgarian language phenomena and traditions in Bulgarian linguistics.

The Logic for WordNet provides an uniform, powerful and clean formalism for expressing complex queries and conditions over the WordNet structure. This enables the usage of one and the same back-end realization for completing different tasks on different WordNet levels – syntactical

structure, data completeness, and semantical inconsistencies in the system.

New semantic relations could be easily included in the logic structure. Further direction of development is extension of the logic with other relations like anthonomy, meronymy, etc. An efficient implementation of the decision algorithm is under development.

The verification methodology is formulated and applied to the Bulgarian data - as a result the Bulgarian WordNet is complete and consistent according to the requirements and the specifications defined in the BalkaNet project.

References

[1] P. Blackburn and J. Seligman, Hybrid languages, Journal of Logic, Language and Information, 3(4), 2, 1996, 51–272.

[2] Cruse, D. A. Lexical Semantics. Cambridge: Cambridge University Press, 1986.

[3] Fellbaum C. (ed.). WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press, 1998.

[4] Koeva S. Bulgarian WordNet - Bulgarian Studies, Ohio, 2003.

[5] Miller G. A. Introduction to WordNet: An On-Line Lexical Database. In "International Journal of Lexicography", Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. Vol. 3, No. 4, 1990, 235–244..

[6] S. Passy and T. Tinchev, An Essay in Combinatory Dynamic Logic, Information and Computation, vol. 93, no. 2, 1991, 263–332.

[7] M. de Rijke, The modal logic of inequality, Journal of Symbolic Logic, 57, 1992, 566–584.

[8] Tinchev T., S. Mihov, S. Koeva and A. Genov, Logic for WordNet, Annuaire of Sofia University, vol. 95, 2003.

[9] Vossen P. (ed.) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer Academic Publishers, Dordrecht. 1999.