# Semantic Network of the UNL Dictionary of Concepts

Viacheslav Dikonov
IITP RAS
Moscow, Russia
dikonov@iitp.ru

Igor Boguslavsky
UPM/IITP RAS
Madrid/Moscow
bogus@iitp.ru

**Abstract**

The article presents a dictionary of concepts, developed as a part of the ongoing effort to create an intermediary language for global information exchange based on semantic structures. The article describes basic principles and contents of the dictionary, which has a potential to become a publicly available language-neutral resource.

## 1. Introduction

This article is dedicated to the development of a new linguistic resource – the Universal Dictionary of Concepts, also known as the Dictionary of UNL (further in this paper – UNLDC). It is a part of a broader international effort to develop a semantic intermediary language named the Universal Networking Language (UNL) [8, 2, 3]. UNLDC is going to serve as the authoritative and exhaustive lexicon of that language. Although the dictionary is closely associated with the UNL language, it has considerable value of its own and can be used as a standalone resource for different scientific and practical tasks not related with UNL.

## 2. Universal Dictionary of Concepts

The basic unit of the UNL language and UNLDC is concept — an abstract semantic unit, coinciding with word senses commonly distinguished by explanatory dictionaries. For example: according to the Merriam-Webster, Collins Cobuild, Oxford and other dictionaries of the English language the word *baby* can be used to express the following five concepts:

*a human child,*
*a cub of a mammal animal,*
*an attractive girl,*
*a childish person,*
*a favorite thing, idea or project.*

Each of them is a separate lexical unit in UNL and has a unique identifier (UW). Normally, there should be only one UW per concept.

The dictionary does not tolerate homonymy, i.e. when one UW is used to express several different concepts.

All concepts are derived from natural languages. The existence of each concept must be supported by some lexicographic evidence of a natural language or a practical necessity, e.g. to express an abstract grammatical meaning or to introduce a non-terminal symbol to organize concepts in the dictionary.

The Universal Dictionary of concepts strives to include and integrate conceptual lexicons of all natural languages. If the dictionary lacks a concept, a new UW should be created on demand and linked with other UWs. It can also be noted that each concept has a definite semantic argument frame of its own.

## 3. Structure of the dictionary

The Universal Dictionary of Concepts must include three principal components:

1. the repository of concepts, commonly referred to as the dictionary of UNL;
2. the network of relations between concepts, which can be referred to as the UNL

Knowledge Base (UNLKB)[1];

3. the local dictionaries, which link concepts with words of various natural languages.

## 3.1 Inventory of concepts

The inventory of concepts is a collection of all concepts available in the dictionary and the UNL language in the form of a flat list of UWs. There is no distinction between UWs for concepts coming from different languages. All concepts are equal as separate lexical units of UNL and listed together. At the same time the dictionary makes it possible to determine the original source language of every concept and all languages that have a direct equivalent.

In principle one concept should be represented by only one UW. However, it is hardly possible to avoid a situation when several different UWs for the same concept appear. It may happen due to technical and organizational reasons in a decentralized community and the dictionary must provide adequate means to handle this situation.

The first and easiest case is when an already existing UW is modified in order to correct an error, achieve better disambiguation or supply missing information. The old version of the UW cannot be deleted immediately, because it can be used by existing UNL documents (or linked to by other resources). Simple deletion would render such documents incompatible with the dictionary. The dictionary has to support per-UW history of changes, allowing to trace any registered version of the UW and prevent reintroduction of deprecated UWs in the same version of the dictionary.

The second source of different UWs for the same concept is the very nature of human language and categorization processes. Each natural language contains a certain amount of exact synonyms e.g. *everyone* and *everybody* in English, which may or may not drift apart with time. It is extremely difficult to build a definitive list of them. Therefore, people will keep adding multiple UWs based on such words even if the corresponding concept already has an UW.

Both processes effectively create groups of UWs resembling synsets used by the Wordnet family of dictionaries. Such groups could be distinguished among all synonyms, viewed as close yet different concepts.

## 3.2 Network of concepts

The concepts create a semantic network linked by the relations of hypernymy, meronymy, instantiation, synonymy, antonymy, association and various other relations describing argument frames. The goal of the semantic network is to provide description of the links between concepts, that exist in the human languages and minds, and make it as objective as possible.

The network of concepts consists of three separate structures formed by a) the ontological relations, which organize the concepts into different semantic classes, b) semantic relations, which reflect similarity or contrast between concepts, and c) argument relations, which specify what argument slots each concept has and the most general ontological classes uniting concepts, which can fill those slots.

### 3.2.1 Ontological structure

The “ontological” structure consists of the UNL relations *icl* (hypernymy) and *iof* (instantiation). These two relations do not exhaust the list of ontological relations, but they have a privileged status. It is obligatory for every UW to specify at least one more general ontological class through these relations. A concept should be linked to all classes, an immediate member of which the concept is. The result is a hierarchy of ontological relations embedded into a network of other relations. Hypernymic classes are hierarchical by nature and with certain approximation can be arranged in the form of a tree, although the real relations between them can be more complex (see figure 5).

1 The semantic network described here is not based on the UNLKB data previously available from [8].
In older UNL publications UNLKB can be referred to as the Master entries dictionary. This name is related with the idea of Master Definitions of UWs – an extended form of UWs, which contains full set of relations with any other concepts. Currently the master definitions are not used, but they can easily be

UNLDC offers a more robust and realistic way to represent the relations between classes of concepts than a regular tree. The resulting base structure is a hybrid one also known as poly-hierarchy. It combines features of a tree and a network. The branches may split and later join, as shown in figure 1, yet there is a common root.

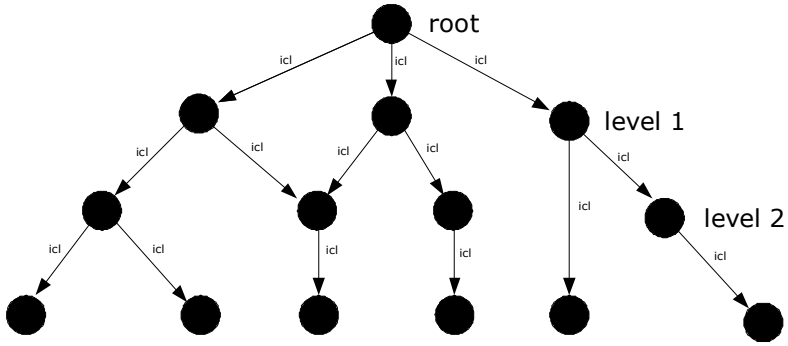


Fig. 1 Ontological structure

The abstract root class is named "uw" (any universal word) and divided into further abstract classes of objects, attributes, actions, states, etc. It is possible to talk about different levels of the ontological structure, but a concept in UNLDC may belong to more than one level or branch.

Ontological relations make it possible to trace the relative semantic volumes of concepts and find more general terms if no direct translation is possible into the target language. For example: while translating the Russian word *zhenit'sya*, which means literally "to acquire a wife" and has no exact equivalent in English, we should replace it with the more general concept "to become married", which has a straightforward English translation "*to marry*".

### 3.2.2 Semantic structure

The "semantic" structure is also ontological by nature, but has a different layout. It consists of the relations *pof* (meronymy), *equ* (synonymy), *ant* (antonymy), *com* (association), which can be supplemented with *fld* (domain of). The *equ* relation does not distinguish between real and quasi-synonyms. Therefore it can be supplemented with other technical means to mark sets of UWs denoting exactly the same concept. The semantic relations unite groups of concepts and do not form any hierarchy. The resulting structure is a pure decentralized network, as shown in figure 2. The dashed line circles in the picture represent groups of concepts with very high similarity.

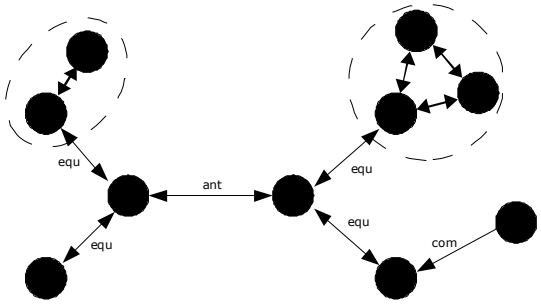


Fig. 2 A fragment of semantic structure

There is no requirement for the this structure to be connected, unlike the ontological one. It may consist of multiple isolated fragments.

### 3.2.3 Argument structure

The argument structure is a collection of argument relations, e.g. *agt* (agent), *obj* (object), *ptn* (partner), *ben* (beneficiary), *plt* (target place), *src* (source), *gol* (resulting state), etc. connecting each concept with an argument frame and general class concepts, which unite all specific concepts that normally fill respective argument slots. In most cases the argument relations point to concepts which belong to a relatively compact group of the most general ontological classes, which occupy

the topmost levels of the ontological structure (Figure 3).

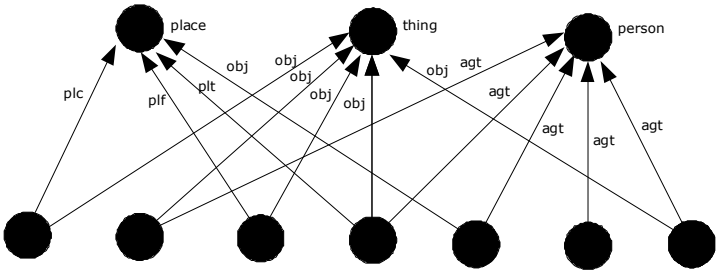


Fig. 3 Argument structure

All three structures link the same concepts and are superimposed on each other, forming the network of concepts of UNLDC.

### 3.3 Local dictionaries

Local dictionaries are used to connect concepts with the vocabularies of different natural languages. Each language should have a local dictionary in order to be supported. The local dictionaries can be just flat lists enumerating pairs of concepts and their translations into the target language. The natural language words may be supplied with grammatic information and morphology information and any other useful data.

A translation does not have to be one word. Some concepts represented by a single word in one language may be translated into another by multiword phrases and abbreviations, e.g. *senior pupil* or *VIP*.

However, not all concepts can be translated into all languages even descriptively. If there is a need to translate such a concept, a nearest general term or a more specific one can be found via the network of concepts. Figure 4 provides an example. It outlines relations between Russian (left) and Bulgarian (right) words for *pen*, *handle*, *knob*, *stem* and *tiller* with UWs as a pivot. There is no direct equivalent in Russian for the Bulgarian word *drzhka* in the sense of stem of a plant. The translation must be chosen by tracing the ontological (icl) links between stem of a fruit and stem of a flower. Additionally, there are two alternative Bulgarian translations for the concept *pen*.

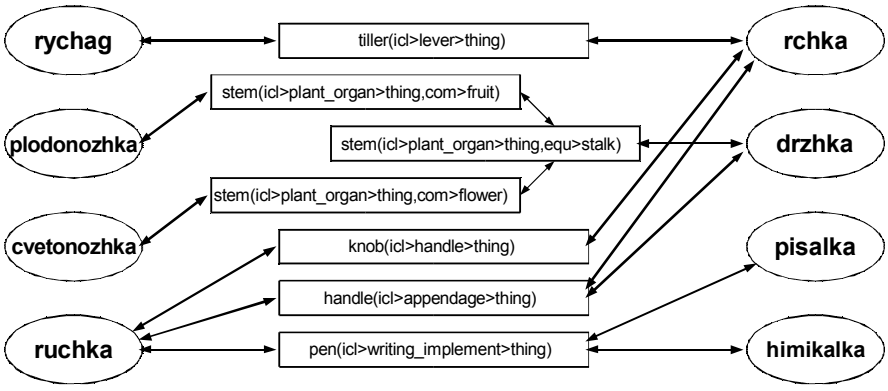


Fig. 4 Concepts and possible links between some Russian and Bulgarian words

## 4. Universal Dictionary of Concepts and Wordnet

The Universal Dictionary of Concepts is quite similar to the well known Wordnet family of dictionaries in many important aspects. Both have concepts as their basic units and define similar relations between them. At the time of writing a lot of data have been imported from Princeton Wordnet [1] automatically. Even more information, including new concepts and relations [5], can be imported from different existing Wordnets into the Universal Dictionary of Concepts. However, there are some important differences between UNLDC and Wordnets.

## 4.1 Relation to natural languages

Each Wordnet describes the lexical system of a particular language and each language is maintained separately. Wordnets may be interconnected by means of the Inter-Language-Indexes (ILI), which describe the relations between the concepts of certain versions of the original Princeton Wordnet (typically 1.5 or 1.6) and concepts of other national Wordnets. However ILIs play a subsidiary role. Only some non-English Wordnets are linked to the original Princeton Wordnet and such links get outdated as soon as a new version of it is released.

The Universal Dictionary of Concepts can be compared to several Wordnets linked through ILI, but ILI link only certain pairs of languages and in most cases one of them is English. UNLDC has no bias towards any particular language. The emphasis is given to the unified inventory of concepts and their relations. Links to vocabularies of natural languages are provided through optional local dictionaries and do not have to be discarded when unrelated changes are made in the repository of concepts and the semantic network.

UWs consist of two parts: a headword and a set of constraints. The headword is usually an English word. However it does not mean that the dictionary uses English as a pivot or a "gold standard" when describing unique concepts of other languages. English headwords and constraints were chosen for mere practicality because it is the only language in common for all participants of the UNL project. If a foreign concept has no exact equivalent in English, it is possible to use constraints to produce a new unique UW from another UW with an English headword. It is also a fact that not all UW headwords are English.

Non-English concepts may also be used as a base for modification and as constraints to describe other concepts. For example:

*samovar(icl>boiler>concrete_thing,com>tea)*
*tula_samovar(icl>samovar>concrete_thing,com>tula(iof>city))*
*sauna(icl>sweating_room>place,com>finnish,com>dry)*
*parilka(icl>sweating_room>place,com>russian,com>steam)*
*venik(icl>massage_tool>...com>parilka(icl>sweating_room))*

If the number of concepts unique to other languages increases, the statement about the special role of English in UNLDC will lose ground.

## 4.2 Hierarchical structures

Wordnets organize the noun and verbal concepts into hypero-hyponymic hierarchies represented as trees. Such structures are easy to search and analyze, but pure tree classification does not support partially intersecting classes. Tree structures work well only for the few top classes of a linguistic ontology. For example, Princeton Wordnet has concepts of (tennis) *racket*, and (hockey) *puck* as well as a class for *"sports implements"*. However, *racket* is a member of the class of sports implements and *puck* is not. Instead it is a member of the class of *"disk objects"*. Moving puck to the *"sports implements"* class in a pure tree would cause losing information that it is a disk.

UNLDC is able and strives to accommodate a different less formally hierarchical approach. The basic ontological structure is a network graph which has only some features of a tree. It is normal to have multiple parents to the same daughter node. It allows for more complex relations and more fine-grained classification. Every concept should be linked to all possible immediate hypernyms. For example, the word *sushi* in Wordnet is a direct daughter of the concept *dish* (food). Suppose that we want to introduce further ontological divisions by nationality (sushi is a Japanese dish) and primary ingredient (sushi is made of fish). It is not possible to decide which of the two classes has to be placed higher in the hierarchy, because these classes specify intersecting sets of concepts (Figure 5)[2].

2 Princeton Wordnet provides a way to include a synset into several classes at the same level of its hierarchy too, but this is not common. For example, *key* in the sense of "*a kilogram of a narcotic drug*" is described as both "*a mass unit*" and "*a metric unit*" at the same level and this split is immediately joined at the next level under the "*units of measurement*" class.

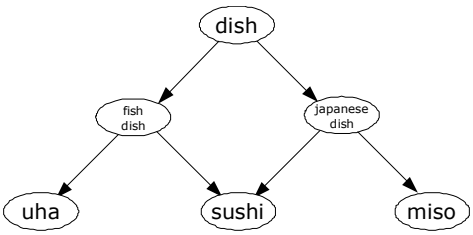

Fig. 5 Multiple parent classes

Using a network instead of a tree has some implications. A tree structure, like that of the Wordnet, allows to trace every concept to its deepest root classes with full confidence, whereas the poly-hierarchy structure permits multiple paths, leading to different and even mutually excluding high-level classes for the same concept. It may cause confusion and disorder. For example, the class *"functional thing",* which includes the concept of hammer, is a daughter of both *"abstract thing"* and *"concrete thing"*, thus making hammer a possibly non physical object! This problem can be remedied in UWs by providing a secondary direct link to the relevant top class.

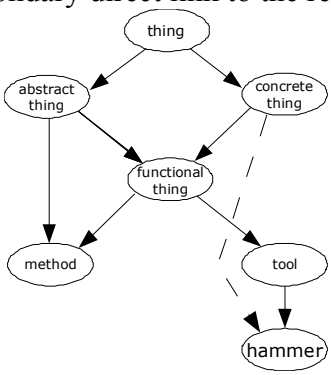

Fig. 6 Additional link to the relevant top class

According to figure 6, the UW for the concept *hammer* should be *hammer(icl>tool>concrete_thing)*. Knowing two ends allows to trace the ontological relations between any concept and the relevant top class and produce optimal single hierarchy from a poly-hierarchy structure.

## 4.3 Other features

Unlike Wordnet the Universal Dictionary of Concepts does not limit itself by certain parts of speech. It provides full set of concepts for prepositions, conjunctions and some words with special grammatical functions, e.g. modal verbs. The UNL language does not maintain POS distinctions and does not limit the range of meanings that can be expressed by UWs.

UNLDC provides more detailed semantic frame information, not limited to the verbal concepts. All argument frame slots are marked by UNL relations. Additionally the most general ontological class suitable to fill each argument slot is specified. Wordnet-like resources sometimes provide information about typical context of synset members, but there is no common approach. Princeton Wordnet shows example sentence frames but has no semantic classification of the types of argument except that it calls some of them "somebody" and other "something". However the yet unpublished dictionary Russnet [7], which is the most promising Wordnet project for the Russian language, is going to have good description of verbal argument frames [6].

Some wordnets preserve syntactic information about the words, such as part of speech, gender, animacy, etc. [9], while other are coupled with morphology engines. This is not the case in the UNLDC because such information is unneeded for the UNL language. The proper place for such data about words of various natural languages is in the local dictionaries.

## 5. Development of the dictionary

The development process should follow the essential principles of division of labor, gradual development, reuse of existing data and decentralization. An open community model is the best option, because no single authority can have enough resources and expertize to do everything.

Every time when a significant amount of changes is done and no formal objections received, a snapshot of the dictionary should be taken and released as a new version. From that moment all participating parties must update their tools to use the new dictionary.

The Universal Dictionary is going to be released to the public under a free license as soon as the first version will be ready, which presupposes merging in more UWs from other UNL groups and putting in operation the infrastructure for automated data exchange. The data may be used freely for any purpose, though commercial use may be a subject to special conditions. Everyone will be given the right to expand the resource and fix errors, provided that all modifications will be returned to the community of dictionary users and editors. The quality of data submitted to the dictionary must be assessed by experts.

## References


[1].Bekios J., Boguslavsky I., Cardeñosa J., Gallardo C. *An Efficient Method for Building Multilingual Lexical Resources* // Proceedings of the Fifth International Conference Information Research and Applications i.TECH 2007, T.1. Sofia.: 2007.C.39-45.

[2] Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L. *The UNL Initiative: An Overview* // Computational Linguistics and Intelligent Text Processing.2005.

[3] Boguslavsky I.M., Dikonov V.G. *Universal Dictionary of Concepts* // Proceedings of the first MONDILEX workshop “Lexicographic Tools and Techniques”. M.: 2008. C.31-42.

[4] *WordNet: An Electronic Lexical Database* Ed. By Fellbaum, C. // MIT Press.1998.

[5] Iraola L. *Using WordNet for linking UWs to the UNL UW* // International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies.Alexandria.: 2003

[6] Azarova I.V. *Shemy upravleniya I ramki valentnostej v RussNet* // http://project.phil.pu.ru/RussNet. 2005

[7] Azarova I. V., Mitrofanova O. A., Sinopalnikova A. A. *Komp'uternyj tezaurus russkogo yazyka tipa WordNet* // Dialog 2003. M.: 2003.

[8] Web site of the UNL project http://www.undl.org.

[9] Suhonogov A.M., Jablonskij S.A. *Razrabotka russkogo WordNet* // RCDL2004.Puschino.: 2004.