Universal Dictionary of Concepts

Viacheslav Dikonov IITP RAS (ИППИ РАН) sdiconov@mail.ru

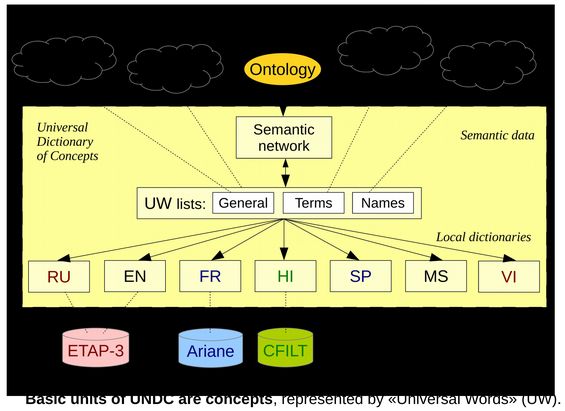
Background and Goals

The initial motivation of the “Universal Dictionary Of Concepts” (UNLDC) was to provide the common standard lexicon for the artificial interlingua UNL (Universal Network Language) and unite multiple existing UNL dictionaries. It follows standards set by the U++ version of UNL.

UNLDC can be used in non-UNL projects as:

- semantic pivot dictionary
- set of lexical bindings for SUMO ontology
- dataset for generation translation dictionaries for rare language pairs (Dikonov 2009)
- tool to create Wordnets (e.g. Yet Another Russnet uses data from UNLDC)
- sense inventory for corpus annotation

Structure of UNLDC dictionary



Basic units of UNDC are concepts, represented by «Universal Words» (UW).

- Each UW consists of a headword and a list of restrictions, which are used to narrow down the semantics of the headword and resolve its ambiguity.

  headword (**icl**>hypernym>class, **equ**>synonym, **agt**>class, **obj**>class)

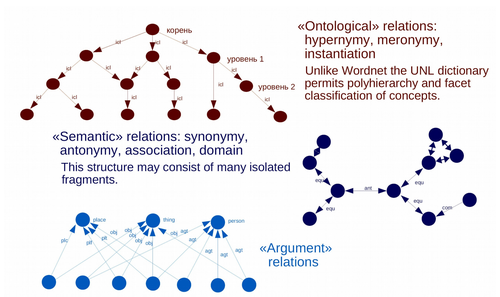
- Types of restrictions:
  - **Ontological** codify the general knowledge about the world: icl (inclusion into a class), pof (part of), iof (instance of).

    tongue(**icl**>concrete_thing,**pof**>body)
  - **Semantic** help to distinguish between concepts that have one common headword: equ (equivalent), ant (antonym), com (component), fld (domain).

    ably(icl>how,**equ**>competently,**ant**>incompetently,**com**>skill)
  - **Argument** reflect the typical argument frame in terms of UNL relations: agt (agent), cag (co-agent), obj (object), plc (place), tim (time), rsn (reason)...

    buy(icl>get>do,**agt**>person,**obj**>thing,**cob**>thing,**src**>thing)
- According to the U++ standard, each UW should have only the minimal set of restrictions necessary to express the difference between concepts with other UWs with the same headword.

Semantic network



- **Work in progress** - Partially available online (see file «links-unl-uw.csv»)
- Semantic relations (96125 relations coming from 44509 UWs)
  - Contains synonymy/antonymy relations from PWN.
  - Over 5000 additional relations based on disambiguated lexical derivation links from PWN.

    a run «the act of running» means the **same** process as *to run «move fast by using one's feet»*
    a walker «a person who travels by foot» is an **agent** of *to walk «use one's feet to advance»*
    a sound «mechanical vibrations...» is an **instrument** of *to sound «announce by means of a sound»*
  - Part-Whole relations with additional annotation of optional and detachable parts.
- Ontological relations (Only a small number of UW↔UW links pulished, but UW↔SUMO links available)
  - Based on an ontology, which extends SUMO with Wordnet hypernymy relations.

    33868 classes and 19804 individuals.
  - Main problem is poor / inconsistent classification of concepts.
    - Refining of UW↔SUMO links is underway and has advanced a lot!
- Argument relations (3850 predicate UWs have manually assigned argument links)
  - Other UWs have argument frames predicted by a statistical tagger based on syntactic government patterns and semantic classes (Dikonov, 2012) but this data needs review.

Statistical overview

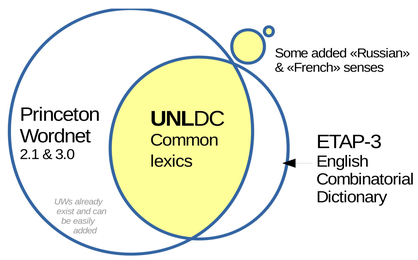
Part	Number of UWs	Status
General lexics	90720	Downloadable from github
Terminology	688617	Prototype exists
Named Entities	2109240	Available, needs an update

- The main priority is given to the general module of UNLDC, which covers most frequently used words of English and Russian.
- Named Entities module is a conversion of DBPedia into the UNLDC format + random additions, e.g. from PWN.
  - Some NEs are still present in the general module because they are homonymous with common English words.
- Terminology module contains domain terms registered in multiple terminological dictionaries and/or domain ontologies. Its taxonomic classification part needs a lot of attention.

Lexical volumes

Language	General lexics	Terms	Names	Total UWs	Quality estimation
English	**87607** (42134 words)	688617	2109240	**2 885 464**	*****
Russian	**61323** (37999 words)	688613	226595	**976 531**	**** Manual proofreading
French	**37012** (25620 words)	103060	367888	**507 960**	*** Automatic verification
Hindi	**42412** (38661 words)	0	10823	**53 235**	**** Updated with Hindi WN data
Spanish	**11741** (6983 words)	21990	298674	**332 405**	** Experimental
Malay	**21847** (17457 words)	0	46044	**67 891**	** Experimental
Vietnamese	**7058** (6704 words)	0	171367	**178 425**	*** Experimental

- English local dictionary is a subset of the Princeton Wordnet, extented with functional words and some terms found to be missing, e.g. certain phrasal verbs.
- Russian dictionary has mixed origin partly generated automatically and partly manual. It is being proofread and extended.
- Hindi dictionary combines data of the Hindi UNL dictionary and Hindi Wordnet.
- Other local dictionaries are products of automatic mapping of dictionaries built by other UNL groups or experimental word sense detection based on lexical networks built from conventional bilingual dictionaries.



Links to external resources

Part	Number of UWs	Connected with
General	81996 (out of 90720) 33343 (out of 90720) 85431 (out of 90720)	Princeton Wordnet 2.1 и 3.0 Hindi Wordnet SUMO ontology
Terminology	all part	Upper SUMO classes (not reliable) Domain ontologies
Named entities	all	DbPedia ontology Upper SUMO classes

- The ontology showed in orange in the dictionary structure diagram is an extended custom OWL version of the Suggested Upper merged Ontology (SUMO). There is a tool that rebuilds it from any new version of SUMO, re-applies all extra modifications.
- Links to Princeton Wordnet cover versions 2.1 and 3.0. They follow the formats specified in the Wordnet documentation. URI references should be added as well in the next versions.
- Possible new additions to the list of linked resources: OpenCYC, Verbnet and/or FrameNet, national wordnets (Russnet? YARN?). Other open semantic resources are welcome.
- English, Russian and French words and phrases have extra references to the dictionaries of MT systems, which are supposed to be able to convert between these languages and UNL.

Characteristic features

UNLDC contains certain concepts that are not directly related to lexical senses of natural language words.

- Abstract concepts known as Lexical Functions (LF)
  “Collocate” type LFs may be used to avoid faulty literal interpretation of certain idiomatically used words.

  Example: words *take* and *have* in *I took a short walk* and *I had a short rest* can be replaced by a special UW *perform_an_action(icl>do,agt>thing,obj>process)* corresponding to the LF OPER1. Its translation into other languages depends on the LF argument.
- Modal predicates
  In many languages, including English and Russian, modal words have different meanings in different contexts. In UNL the same modal attribute or UW is used to encode a given modality regardless of what modal word was used in the source sentence.

  The prohibition in *You may not carry a weapon here* and *You can not smoke onboard* is expressed by two different modal verbs but in UNL the same symbol is used to represent both. The dictionary will link the special modal UW *grant-not(icl>modal>be,obj>uw,aoj>thing)* to both *can not* and *may not*, leaving the choice between them to the processor or the user.
- Abstract ontological concepts

  PhysicalObject, ContentBearingObject, IntentionalProcess ...

Full and quasi synonymy

- Regular synonymy relation between members of Wordnet synsets is viewed as quasi-synonymy and marked with the UNL *equ relation.*
- UWs created due to spelling variations of the same word in English are linked with *cnt* relation and can be merged together.
- Some UWs have several translations into the same natural language. Such translations are supposed to be full synonyms too.

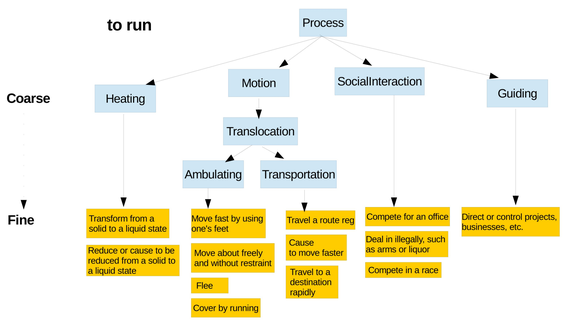
  hydroplane(icl>airplane>thing,equ>seaplane) is linked to two Russian words *гидроплан* and *гидросамолет* with exactly the same meaning.

Multiple level of semantic grain

1) Individual concepts are the smallest possible grain level.
   Imported synset members are treated as independent quasi-synonyms.
   * Translations to other natural languages may be approximate and blur fine distinctions.
2) Synsets
3) Taxonomy classes. Different senses of the same word often fall into the same or related ontological classes
   1) *to run* in the senses *«melt something»* and *«become melted»* falls into the class *Heating*
   2) *to run* has 11 senses falling into subclasses of *Motion «move fast by using feet», «flee», «travel a route», «cover by running», «cause an animal to move faster», «pusue for food or sports»…*
      These further subdivide into: *Ambulating (Running & Walking) 4 senses, Translocation 4 senses, etc.*

      3 senses *«compete for an office», «compete in a race», «deal in illegally»* fall under the class *SocialInteraction*

      3 senses *«have a specified tendency» «make without a miss» «be subjected to»* fall into subclasses of *Attribute*



Multiword phrase structure

Concepts may be translated into some languages by multiword phrases and/or grammar constructions with a variable part. The target language expression may be written not only in the form of a simple n-gram, but also include syntactic relations between the words following the notation of the MT system that is supposed to support the target language.

bathroom(icl>room>thing) is translated into French as *salle de bain*. The phrase structure recorded in the format of the French MT system Ariane looks like:

@@_1:'salle' (2:'bain')::1 '': CAT(CATN),GNR(FEM),N(NC). 2 '':CAT(CATN),GNR(MAS), N(NC), NUM(PLU),ART(ABS), RSUNL(MOD)

A more generic option is to establish UNL relations between words of a phrase. UNL-capable systems should be able to convert such notation into their internal syntactic structures.

salle – pur→bain «hall for bathing»

UNLDC and Wordnets

- Universal Dictionary of Concepts has a lot in common with Wordnets

  Princeton Wordnet is the source of many UWs and the English local dictionary Concepts derived from the Wordnet have back references to PWN 2.1 and 3.0
- **Differences** between UNLDC and Wordnets:
  - The basic units are concepts/UWs
    - By default members of synsets are treated as **quasi**-synonyms, which may have subtle differences, e.g. sentiment.
    - Synsets exist as groups of concepts linked with the synonymy relations (equ, cnt).
  - No intentional bias towards any single natural language
    - non-English concepts (currently 4,6%)
    - LF, modals, abstract ontological concepts
  - No separation between parts of speech
    - Includes prepositions and conjunctions
  - Different organization of the semantic network
    - polyhierarchy instead of tree structure
    - top levels are based on a formal ontology (SUMO with extensions)
    - argument structure

How is it built?

- Method: accumulation and integration of available data with subsequent proofreading and gap filling.
  - Regular translation dictionaries and Wordnets were primary data sources for NL lexicons.
  - Most of the links between UWs and natural languages were first built automatically.

    Quality assessment is extremely important !
- All UW↔NL word links have tags that group them according to their expected quality, manual or automated way of creation, number of senses initially ascribed to the word, etc.
  - Autogenerated links are subject to automatic ranking based on the number of sources confirming translations, that can be deduced from the UNL dictionary.

    Example tags: manual, auto, auto-good, auto-monosemic, auto-polysemic-3lang…
  - Each tag gives numerical points to a link, which can be summed up and normalized to roughly estimate the expected quality level.

    English volume rating: 100%

    Russian volume rating: 96,7% confirmed

    Vietnamese volume rating: 47,7% confirmed

    French volume rating: 16,44% confirmed
- Russian dictionary entries were ranked using distributional semantics methods (Dikonov & Poritsky, 2014).

Where are the data?

https://github.com/dikonov/Universal-Dictionary-of-Concepts

Old snapshots: http://atoum.imag.fr/geta/User/services/pivax/data/

Data are free to use under GPLv3+ / CC-BY-SA / CC-BY-SA-NC

If when you find an error, you can send a notice or corrections to sdiconov@mail.ru.

Data formats:

- CSV Simple TAB-separated text tables in Unicode UTF-8 encoding.
- XML (pivax, xdxf)

  XML formats used to import data into various dictionary shells or Papillon/Pivax lexical databases.

  There are desktop dictionary shells (stardict, goldendict, dicto...) available for every OS that support XDXF XML directly or via conversion to their native formats. They nay be used as a convenient way to query the data.
- LMF / RDF/Turtle (Open Linked Data)

  Standardized distribution formats supported by a large community

  To be added.

Additional data

- Russian synsets - A list of Russian synonyms automatically derived from UNLDC. Updated when changes are made to the Russian volume.
- Russian-Hindi and Hindi-Russian translation dictionaries - Generated automatically using semantic pivot. The same can be done for any other language pair.

Other related projects

- Converter of English and Russian text into UNL graphs
  - Multilingual pivot translation
- Ontology based semantic parser for Russian
  - Logical inference

  Both use ETAP-3 platform.