Establishing Links between Natural Languages and the Universal Dictionary of Concepts

Viacheslav Dikonov

Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow

Abstract. This article explains how to create dictionaries, which would link the vocabularies of any chosen natural languages with the Universal Dictionary of Concepts [3] and the pivot language UNL. All languages linked with the Universal Dictionary of Concepts become automatically linked with each other at the semantic level of word senses. The article describes the minimal requirements for the contents of such dictionary, explains the principle of data exchange and suggests a possible procedure of producing the dictionaries by merging already existing common lexicographic resources.

1 Introduction

The Universal Dictionary of Concepts (UDC) [3] is the definitive repository of concepts forming the lexicon of the Universal Networking Language (UNL) [4]. The UNL language enables computers to record the meaning of a natural language text, store and exchange semantic information in a standardized form. UNL has many potential applications. For example, it can serve as a pivot language for automatic translation or facilitate unambiguous search in multilingual environments.

There are several linguistic processors developed in different countries, which support the UNL language[1]. Systems which translate text into UNL (enconversion) are called UNL converters. UNL Deconverters are systems that perform the reverse operation (deconversion) and turn UNL documents into texts in some natural language. The list of languages already having a UNL deconverter includes English, Russian, French, Spanish, Arabic, Japanese and more. UNL represents the meaning of a text as a graph joined by semantic relations. The graphs can be visualized and their visual form is intuitively understandable.

The basic elements of UNL and UDC are concepts. Concepts are understood as abstract semantic units more or less equivalent to word senses commonly distinguished by explanatory dictionaries. However, concepts are not bound to concrete words or idiomatic phrases of any particular language. All concepts have their origin in natural languages and should be supported by some linguistic source or a practical need.

Each concept is unambiguously represented by a Universal Word (UW) [2,3,4]. Every UW stands for one and only one concept. Any new concepts receive their own unique UWs. It is possible for technical reasons to have several UWs for one concept (strict synonyms) but such situation is undesirable and should be avoided if possible.

UDC consists of three parts: the repository of concepts, a semantic network establishing relations between concepts, and a number of local dictionaries establishing links between concepts and words or expressions of natural languages. Every language should have its own local dictionary. UDC will be a free public resource constantly developed by the UNL community and any other interested parties.

[1] The projects of making a UNL enconverter and deconverter for the Russian and English languages have received funding from the Russian Foundation for Basic Research (RFBR) under grant agreements 08-06-00367 and 08-06-00344.

2 Local dictionaries

2.1 What is a local dictionary?

Local dictionaries as a whole are one of the key elements of the UNL infrastructure enabling the intermediary language to perform its function of capturing and recording the semantics of any natural language text. Each local dictionary provides a lexical interface between a single natural language and UDC. Any lexicographic resource that describes the polysemy of words of any natural language by linking them with UWs of UDC will qualify as a local dictionary in terms of UDC. Local dictionaries can be used by UNL converters and deconverters to perform automatic or semi-automatic conversion between a natural language text and its semantic representation in UNL.

The exact content of a local dictionary is determined by peculiar properties of the natural language it describes. It is hardly possible to set a rigid standard in this area, but certain common guidelines and principles are essential for interoperability.

A local dictionary can be used for:

1. making the graphical form of the UNL semantic graphs more intuitive for a casual reader or author, who wants to verify the semantic representation of his work
2. semantic markup of corpora, disambiguation of keywords for performing search in UNL or multilingual environment, other cases when lexical disambiguation is necessary
3. finding relations between words of different languages to produce translation dictionaries automatically
4. UNL conversion and deconversion, automatic translation.

Each of the four uses sets different and progressively greater quality and content requirements for a local dictionary. Every new dictionary can be developed gradually through a process of iterative refinement that would make it increasingly bigger, better and more useful. The entry level can be low enough to allow practical use of a bare minimal local dictionary which is just a list of word lemma and UW pairs.

2.2 Levels of quality

The first of the four uses listed earlier is the least demanding. There are specialized software tools to visualize and edit UNL graphs in order to post-correct any errors of an automatic converter. The UWs of UNL are rather long and less familiar to a novice user, so some editors provide an option to display translations instead of the UWs. It helps to see words of a different human language inside the nodes of the graph to quickly assess the quality of lexical disambiguation and spot important errors. Even an incomplete or autogenerated preliminary version of a local dictionary might serve this purpose as soon as it is free from obvious errors. Figure 1 shows an example of a very simple but already useful local dictionary.

Word	**Universal Word**
сказать	say(icl>communicate>do,equ>tell,agt>person,obj>uw,rec>volitional_thing)
сказать	tell(icl>narrate>do,cob>uw,agt>person,obj>uw,rec>person)
сказать	say(icl>order>do,agt>volitional_thing,obj>uw,rec>volitional_thing)
сказать	say(icl>imagine>do,agt>person,obj>uw)
человек	person(icl>abstract_thing,equ>personality)
человек	one(icl>unit>thing)
человек	mankind(icl>homo>thing,equ>world)
человек	human(icl>hominid>thing,equ>homo)

Fig. 1: A fragment of a minimalistic Russian local dictionary

The second goal, i.e semantic markup of corpora, is much more demanding from the point of view of dictionary's coverage, correctness and precision. At the same time, the dictionary can still be a simple list of word-UW pairs, supplemented with definitions and examples. The existence of several local dictionaries in UDC makes it possible to retrieve definitions of the concepts in different languages, as shown in Figure 2. The English local dictionary already contains definitions and examples for all concepts in the current version of UDC and POS classes of the linked words are easily deductible from the UWs[2].

Word	**Universal Word**
человек	man(icl>person,equ>human,ant>animal)
	человеческое существо // отряд в пятьдесят человек
	a human being // a hundred men died
человек	person(icl>abstract_thing,equ>personality)
	совокупность черт характера // приятный человек
	the personality of a human being // a nice person
человек	one(icl>unit>thing)
	всякий, любой человек // человек никогда не должен себя ронять
	any person as representing people in general // one should never be complacent
человек	mankind(icl>homo>thing,equ>world)
	человеческая цивилизация // человек шагнул в космос
	all of the living human inhabitants of the earth // one giant leap for mankind
человек	human(icl>hominid>thing,equ>homo)
	биологический вид // человек умелый
	the genus homo // the evolution of humans
человек	man(icl>subordinate>person,equ>agent,pos>person)
	зависимое лицо // человек Путина
	a male subordinate or agent // our man in Habana

Fig. 2: A fragment of the Russian local dictionary with definitions and examples from two local dictionaries

The third possible goal of matching words of multiple natural languages for automated construction of translation dictionaries represents a whole new level of requirements. A very detailed and precise description of polysemy is needed to establish correct translation pairs. Some additional information, such as pragmatic usage tags, e.g. *poet*, *archaic*, *informal*, good definitions and examples in all matched languages, becomes mandatory. Other types of information typically provided by translation dictionaries include morphological and grammatical features, phonetic transcription, sample sentence structures, etc. Figure 3 shows local dictionary entries containing enough data to fill a typical translation dictionary entry and how they combine.

To achieve good results, the coverage and degree of precision should be comparable for all languages involved and sufficient to establish correct translation pairs.

Finally, the fourth and most important use of a local dictionary is automatic translation (MT) through UNL conversion and deconversion. Different linguistic processors set different standards for their dictionaries. Usually such applications favor generalization of word senses to lessen the complexity of dictionaries and disambiguation procedures employed at the stage of syntactic analysis. On the other hand, automatic translation requires full morphological and grammatical information as well as knowledge about combinatorial potential of the word.

[2] All UWs have specific descriptors corresponding to parts of speech provided by the *icl* relation: *do, be, occur* – verbs, **thing, person, animal* etc. – nouns, *adj* – adjectives, *how* – adverbs, *how* in combination with an *obj* constraint – prepositions.

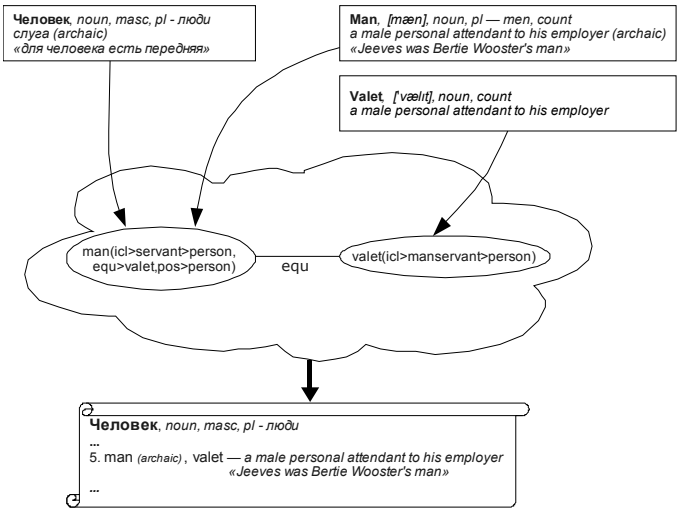

Fig. 3: Russian and English local dictionary entries linked through UDC provide data for construction of a translation dictionary

2.3 Data Exchange

Since local dictionaries are optional parts of UDC and most of them are going to be maintained separately by independent teams, there will be no technical requirements for the storage format or a prescribed set of tools. Instead, there will be a requirement to maintain compatibility of data with UDC and ensure regular reciprocal data exchange. It means that all local dictionaries must synchronize with each new release of UDC to accommodate to any changes in the UW set. At the same time, any changes in a local dictionary that result in adding new concepts or changes of relations between concepts must be submitted to UDC.

Each local dictionary must be machine readable. UDC is going to be stored in an SQL database table, so the local dictionaries should be ready to export and import data in Unicode in a compatible table form either as CSV or XML. The exact technical description of the exchange format does not exist yet. It is going to be designed together with the Internet infrastructure for the UNL dictionary following the availability of the first public release.

All local dictionaries must export at least one data field containing lemmas of the words or expressions associated with UWs of UDC. This field and any additional fields with extra kinds of data are called public. All public data fields involved into the data exchange process need to be marked in a standard and consistent way across all local dictionaries, but their contents may be language specific. A dictionary may contain certain data not relevant to the UNL and UDC project or excluded from the data exchange. Such fields are called private.

We consider it a good practice to keep a copy of every local dictionary that would include all public data fields in the central public database as a safety and informational measure. It will make editing of UNL graphs more convenient by enabling on-the-fly switch from UWs to words of any desired language and help to rebuild any local dictionary in the event of data loss or if the original team ceases to exist.

3 Making of a local dictionary

3.1 General steps

The process of making a local dictionary includes several steps. Some of them can be automated or significantly simplified by re-using existing lexicographic resources and merging their data. The steps are:

1. Identification of word senses (concepts) of a target natural language and definition assignment.
2. Matching of the word senses of the natural language with existing UWs.
3. Creating new UWs for concepts that could not be matched exactly.
4. Linking the new UWs into the semantic network of UDC. It can be done in parallel with stage 3.

This work is quite similar to creation of a Wordnet for the target language. Languages that already have a Wordnet with a good ILI linking it with recent versions of the Princeton Wordnet will have a substantial advantage. Most of the UWs in the current version of UDC are prepared on the basis of Princeton Wordnet [1] v2.1 and can be traced back to the corresponding synsets. UDC will maintain its links with Wordnet to simplify data migration in both directions. Any new and edited UWs, which have their counterparts in Wordnet, should be included in the UDC-Wordnet list of correspondences. Each concept added to UDC will be tagged with its source language. All concepts will also carry a tag with the list of languages that have an exactly matching word sense. The semantic network of UDC will include all links and hierarchy provided by Wordnet and extend it with any missing relations. The combination of these measures will make it possible to extract a Wordnet-type resource for any linked language from UDC.

3.2 Matching word senses and UWs

The list of word senses and their definitions for a chosen language is usually available in the form of an explanatory dictionary[3] while the list of UWs will be provided by UDC. Each UW already has a supposedly self-explanatory name, a definition in English and sometimes an English example. At the current stage of development there are about 200 000 UWs covering the lexicon of the English language and all of them use English words as headwords. It is possible to use a translation dictionary to find English translations of a word. UWs with headwords matching the English translations of the chosen word create a list of candidate UWs for each word sense.

The next step is matching the word senses of each word with candidate UWs. (Fig.4).

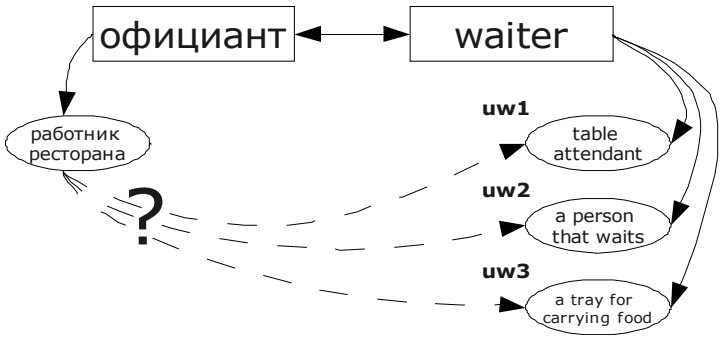

Fig. 4: The word sense matching problem

[3] If no explanatory dictionary is available, as it might happen with some less studied minority languages, there are other ways to identify word senses, e.g. by using text corpora or translation dictionaries.

The choice based on definitions is simple enough for a small number of words but doing it for all UWs is a lot of work. Therefore, it is convenient to have yet another source of information that would help to find certain pairs of word senses automatically. Existing national Wordnets, such as those built for Bulgarian and Czech by the Balkanet project have less coverage than UDC and the Princeton Wordnet, but they provide valuable data for the most frequently used and most polysemous words.

There is a difference between Wordnets and UDC, which becomes evident at this stage. UDC does not treat synsets as monolithic atoms of meaning. Each entry of the dictionary is a single UW. UWs and are still joined by the synonymy relation *equ* into synsets, but UDC permits independent modification of synonyms, recognizing the possibility of subtle differences between them. The synonymy relation is understood as a relation between close but not exactly similar units. Therefore, each synset imported from a Wordnet resource and matched with a set of UWs will produce a set of word-UW pairs (see Fig. 5). Such pairs have high probability of being correct but they must be put to scrutiny as well.

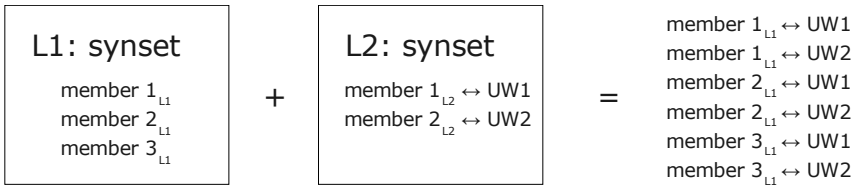

Fig. 5: The result of importing two synsets linked by an ILI

When the process of matching of the word senses with existing UWs is completed, there will be a certain number of word senses left without a matching UW. It is normal, because each language has its own unique conceptual lexicon and it is never fully identical with lexicons of other languages for cultural and historical reasons. The word senses in this list should be added to UDC as new concepts.

3.3 Adding new concepts

Any concept existing in the form of a distinct word sense in any of the linked languages and not found in UDC may and should be added there. A new concept must receive a unique name – a new UW. Local dictionaries cannot reference any UWs not submitted to UDC. Failure to do so may cause incompatibilities between different local dictionaries. There is a standard for UW construction adopted by active UNL centres in Grenoble in 2007[4]. All new UWs submitted to UDC must follow this standard. Malformed UW will be rejected. The designers of the standard can arrange short training courses for those who need to create a large number of UWs.

Every UW consists of a headword and a set of constraints, which describe how the concept represented by the UW is different from the concepts represented by other UWs with the same headword. A constraint consists of a UNL relation and another UW, usually reduced to its headword. The general UW format is:

headword(relation>uw>uw,relation>uw,...)

The headword is usually an English word or phrase. New UWs for concepts related with some previously known concept must be derived from an existing UW by adding or changing constraints. The new constraints must reflect the difference between the new concept and the old one. For example, the first of the following three UWs stands for a general concept of entering into a marriage. The other two are its hyponyms describing two aspects of the action differentiated by some languages, including Russian.

[4] The full description of the standard and detailed guidelines for constructing new UWs are described in a special manual [2]. The manual is still being updated in parallel with the refinement on the initial set of UWs. This work should be completed in summer 2009.

marry(icl>do,agt>person,obj>person) "заключать брак"
*marry(icl>do,agt>*man*,obj>*woman*) "жениться"*
*marry(icl>do,agt>*woman*,obj>*man*) "выходить замуж"*

If the new concept is culture-specific and has no hypernym in English, we can use the native word transliterated into Latin and supplement it with constraints that would link it with the nearest commonly known class of objects.

tarator*(icl>soup(icl>food)>matter)*
lapot*(icl>footwear>..,equ>bast_sandal,com>russian_peasantry)*

UW constraints convey only a minimal amount of information required for identification of concepts. There are three types of constraints: ontological, semantic and argument.

Ontological constraints reflect the most important links between concepts: hypernymy (icl), meronymy (pof), instantiation (iof).

*tongue(**icl**>concrete_thing,pof>body) , madrid(**iof**>city)*

Semantic constraints are used to show the difference between several concepts associated with one headword: synonymy (equ), antonymy (ant), association (com).

*ably(icl>how,**equ**>competently,**ant**>incompetently,com>able)*

Argument constraints reflect the semantic frame of the concept: agent (agt), object (obj), second object (cob), source (src) ...

*buy(icl>get>do,agt>person,**obj**>thing,**cob**>thing,**src**>thing)*

More detailed information about the relations between UWs is going to be stored in the semantic network of the Universal Dictionary of Concepts.

3.4 Linking of concepts into the semantic network

All new concepts should be linked into the semantic network of UDC to maintain integrity of the common dictionary. Linking a concept requires answering several questions, which are usually addressed at the time of construction of a new UW:

1. What is an immediate hypernym or hypernyms of the new concept?
2. What are the immediate hyponyms of the concept?
3. Are there any exact synonyms?
4. Are there any antonyms?
5. What is the semantic argument frame of the concept?

It is possible to create a special software tool to add new concepts to UDC that would provide a wizard interface and reference information to guide the user through the process of creating a new UW and linking it.

3.5 Why linking of new concepts is important

Linking of new concepts extends the semantic network component [3] of UDC. One of its functions is to ensure the ability of UDC and UNL to serve as a pivot for multilingual translation. UDC must always provide a way to find some translation for any word of any supported natural language into any other supported language.

However, objective differences between languages and different approaches towards the degree of granularity and precision of definitions taken by lexicographers will cause situations when different

languages will link to closely related yet different UWs. While Princeton Wordnet sets a common standard it is not always consistent in this aspect. It may happen that some local dictionaries, especially the ones based on richer source data, will go into greater semantic detail while others will link to more general concepts. As a result, some translation equivalents will never be matched (See Figure 6).

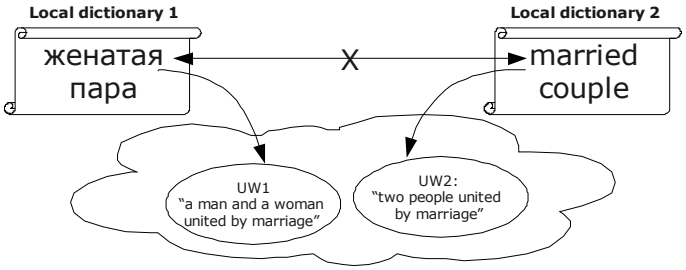

Fig.6: Two words linked to different concepts cannot be matched

A translation for any concept can be found by tracing the ontological (inclusion, instance of, part of) and semantic (synonym of) relations of the semantic network. The rules of finding a translation for a concept that lacks a direct translation into the desired language can be outlined as follows:

1. If a synonym of the concept has a direct translation (member of the same synset), take it.
2. If the concept has immediate hyponyms with translations, choose one of the hyponyms by examining the context e.g . to translate *pedicle* as either *цветоножка* (stem of a flower) or *плодоножка* (stem of a fruit). This is only possible for MT systems.
3. Follow the hypernymy chains until the nearest hypernym with translation is found. If there are several possible paths in the web-like structure, take the shortest one leading to the top parent class specified by the icl restriction of the UW.

The general effect of the third rule applied to an incomplete dictionary resembles the casual speech or speech of an uneducated person, e.g. *give me that thing* (because I do not know its proper name).

4 Summary

This article extends the description of the features and structure of the Universal Dictionary of Concepts in [3]. It shows how to make a local dictionary on the basis of existing lexicographic resources. The advocated incremental manner of development and refinement of a local dictionary allows to obtain some practical result from early steps and find new applications when the quality, content and size become sufficient. The proposed data exchange scheme provides maximum flexibility to the dictionary developers by allowing them to link any suitable resources to UDC regardless of the tools and data formats to used maintain them.

The resulting common multilingual dictionary infrastructure can be used for various linguistic purposes not necessarily related with the development of the UNL project itself. The scheme described in this article is designed to avoid resource fragmentation that became a serious problem in the realm of Wordnets, where multiple projects develop without mutual coordination. Absence of a common data repository for Wordnet-like resources causes huge amounts of useless parallel work. A lot of valuable lexical resources became obscure or simply disappeared after being completed for lack of support and technical maintenance. The Universal Dictionary of Concepts offers a chance to change this situation and accumulate lexicographic data in such way that they will always be readily available to researchers.

References

[1] Fellbaum, C. (1998). WordNet: An Electronic Lexical Database, MIT Press
[2] Boguslavsky. I. (2008). Guidelines for UW construction, manuscript
[3] Boguslavsky I., Dikonov V. (2008). Universal Dictionary of Concepts. In *Proceedings of the First open MONDILEX workshop "Lexicographic Tools and Techniques"*, pages 31–55, Moscow
[4] Web site of the UNL project, `http://www.undl.org`