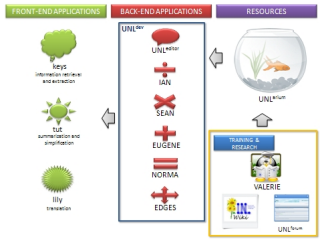
The UNL[web] is a long and often winding road to UNL-based applications. The itinerary may be bedazzling and dizzy, but the travel is always enlightening. And the destination is definitely worthwhile. A tentative map is presented here.



TRAINING & RESEARCH

The UNL[web] is fore and foremost a place to learn and to share. We don't know very much about the structure of natural languages. We do have theories and hypotheses, but most of them seem not to be sufficient or tailored to computational processing. The training & research module of the UNL [web] is dedicated to the collaborative construction of knowledge about UNL and natural languages. It includes three environments: the UNL wiki, a collaborative website used for documentation; the UNL forum, a message board for discussing problems and issues related to the UNL; and VALERIE, the VirtuAl LEarning Environment for UNL, an e-learning facility intended to normalize the vocabulary and level users coming from different traditions and practices in language description. These three environments provide our maps and compasses, the theoretical and

methodological bases of our exploration. They are not only our starting points, but also our travelogue, the constantly updated compilation of our findings in the way.

RESOURCES

Language resources are stored in the UNL arium , the endless repository of what we know about language. The UNL arium is a database management system, where linguists are expected to add, edit and revise dictionary entries and grammar rules. It also contains documents in UNL. The UNL arium consists of several different compartments: lexical databases (dictionaries), rule bases (grammars) and document bases (corpora), which are meant to be bidirectional (i.e., able to be used both in natural language analysis and generation) and as comprehensive as possible. The UNL arium framework assumes that that there can be a single scientific metalanguage able to describe the structures and phenomena related to all natural languages. This metalanguage, consolidated in the UNDL Foundation tagset and in the UNDL Foundation Specifications and Recommendations, has been constantly revised and improved. Attributes, whenever present in a given language, are represented in the same way, with the same tags, with the same formalism, so that all dictionaries and grammars share the same overall structure and have exactly the same syntax. The use of the same standards provides comparable and aligned multilingual databases, and forces the dialogue and the exchange between different traditions and models of language description.

- 

The **UNL Dictionary** (UNL dic) contains the Universal Words (or UWs), the words of UNL, which stand for discrete concepts conveyed by natural languages. In the UNL Dictionary, UWs are listed, defined, exemplified and categorized. The set of UWs is open and subject to permanent increase, but redundancy is to be avoided. For the time being, the UNL Dictionary has been mainly derived from the WordNet3.0, each synset corresponding to a UW. In this sense, the UNL Dictionary reflects mostly the vocabulary of English. But this repertoire is expected to be revised; synsets have been already marked for under-specification or over-specification in relation to other languages; and UWs are supposed to be merged or divided if they prove not to be as comprehensive or as precise as they should be. This is,

however, a perennial work, to be carried along the entire road, as we accumulate knowledge and experience from the other tasks. The UNL Dictionary requires a special certification (CUP [500]).

- 

The **NL Dictionary** (NL^{dic}) contains the words of a given natural language. As in the UNL Dictionary, these entries are listed, defined, exemplified and categorized according to their morphological and syntactic behaviour. The NL Dictionary is also an everlasting enterprise, which is expected to be accomplished in a corpus-driven way, i.e., natural language entries are expected to be addressed as they appear in a given document The NL Dictionary requires a special certification (CLEA [450]).

- 

The **UNLNL Dictionary** contains mappings between UWs and natural language entries. It is a bilingual lexicon where UWs are translated into "lexical realisation units" (lexemes, semi-phrasemes, quasi-phrasemes and phrasemes) of a given natural language, along with the minimal set of features necessary to differentiate between homonyms. As all open-class natural language entries are associated to UWs, the UNL-NL Dictionary also works as the pivot table that allows the alignment of all NL dictionaries.The UNL-NL dictionary requires a special certification (CLEA [250])

- 

The **UNL Knowledge Base** (UNL KB) contains semantic relations between UWs along with a degree of necessity, i.e., the possibility of occurrence. These relations include ontological relations (such as "a kind of", "a part of", "an instance of") - which is to say that the UNL KB encompasses the **UNL Ontology** - and thematic relations (such as "is the agent of", "is the place where", "is the moment when"), which allows for the representation of more detailed information about each entry. The initial status of the UNL KB corresponds to the lexical relations (synonymy, antonymy, hyponymy, hyperonymy, etc) present in the WordNet3.0, but it has been already incremented with the representation, in UNL, of their definitions (the iGLU project). The UNL KB is expected to be dictionary-based, i.e., we intend to represent, in UNL, the definitions presented by several different ordinary dictionaries, from as many languages as possible, in order to cope with the diversity and complexity of each UW. In this sense, the systematic treatment of the definitions from the

WordNet is only the first step into the UNL KB, which we intend to use as the main strategy for word sense disambiguation in natural language processing. The UNL KB requires a special certification (CUP [500]).

- 

The **UNL Example Base** (UNL EB), also known as the **UNL Encyclopaedia**, contains semantic relations between UWs and a degree of probability. However, differently from the UNL KB, which is dictionary-based and focuses on the essential part of the meaning, the UNL Encyclopaedia is corpus-based and covers also the accidental part of the meaning, i.e., information that is extracted from the frequency in the corpus (i.e., related to the probability of occurrence) rather than from the logical structure of the definition (i.e., related to the possibility of occurrence). The UNL Encyclopaedia is also meant to work as a disambiguation device, and is intended to be provided automatically, mainly from the analysis of large corpora. For the time being, this is the only module of the UNL
arium

framework provided by statistical approaches.

- 

The **UNL-NL Memory Base** (UNL-NL MB) contains mappings between natural language structures and UNL graphs and their frequency of occurrence. These mappings are related both to continuous structures (such as n-grams) or to discontinuous structures. They are obtained either by past UNLizations (when they are extracted out of UNL-NL aligned corpora) or by terminological repositories created explicitly by users. They are also used for disambiguation purposes. The UNL MB requires a special certification (CUP
500

).

- 

The **NL grammar** is the set of rules of a given natural language. It is expected to be bidirectional (i.e., to be used both for natural language analysis and generation), and it is divided in five different levels: phonetics, morphology, syntax, semantics and pragmatics. The NL grammar also contains the language settings, which are the set of attributes and values that a language may have. The NL grammar is intended to be done in two different movements: the first, initial, corresponds to the language settings and to the creation of the basic grammar, which is a set of generic (catch-all) rules to be applied in the absence of more

specific rules. This can be done through the Grammar Wizard, which contains a questionnaire of about 350 questions that defines the general behaviour of the language. The second movement is more permanent, and includes inflectional paradigms and subcategorization frames that are created on demand, as required by dictionary entries. The NL grammar requires a special certification (CLEA 700). Only language managers have the privilege of defining the language settings.

- 

The **UNL grammar** is the set of relations and attributes of UNL, which is currently not open for revisions (even though it is expected to be revised in the near future). The current set includes 45 semantic relations and more than 300 attributes, that have been used for natural language analysis and have been object of discussions in the UNL forum. They are documented in the UNL Specs and have been further explained and exemplified in the UNL wiki

.

- 

The UNLarium also contains documents in UNL. The **UNL corpora** is always aligned (UNL-NL) at the sentence level, and has been normally produced by hand, even though machine-aided. It is used mainly for extracting entries for the UNL
eb

and for the UNL
mb

. The UNL corpora is also used to set the standards of UNL and to test the engines and tools that are part of the UNL
dev

, the set of computer systems that operate with UNL.

BACK-END APPLICATIONS

Back-end applications target specialists and require expertise in UNL. They constitute the internal modules of UNL applications and are used to assist linguists in producing more fine-grained UNL-driven resources, such as grammars and knowledge bases. Back-end applications have been released under the UNL
dev

. They may be classified in three different categories: UNLization (i.e., natural language analysis); NLization (i.e., natural language generation); and others.

UNLization

UNLization software take a natural language input and delivers an output in UNL. They are language-independent and have to be parametrized to the natural language input through a dictionary and a grammar, provided as separate interpretable files. For the time being, the UNL^{dev} contains three UNLization software:

The **UNL^{editor}** is a graph-based UNL authoring tool.The whole decision-making process is human: the language specialist uploads the text to be analysed; selects the corresponding UWs (i.e., the nodes in the graph); creates semantic relations between nodes; and assigns attributes to nodes. All is done through a graphic interface that allows users to manipulate high-level graphs instead of low-level UNL statements. The system has been recently improved with a base of past UNLizations that is expected to accelerate this process, but the core decisions remain fully human. The UNL ^{editor} is specially suitable for documents demanding high-quality UNLization, as the ones used in translation.

IAN is the acronym for Interactive ANalyser. Differently from the UNL^{editor}, it includes a grammar for natural language analysis and operates semi-automatically. The word sense disambiguation is still carried out by the language specialist, but the system can filter the candidates using an optional set of disambiguation rules. The syntactic processing is done automatically through the natural language analysis grammar, but syntactic ambiguities are signalled to the user, who may backtrack and choose different syntactic paths. In any case, human interaction is always optional, and is used to improve the results. In case

of no human intervention, the system simply outputs the most likely alternative, which is the one corresponding to the highest priority in the lexicon and in the grammar.

SEAN is the acronym for Shallow Enhanced ANalyser. Differently from IAN, it is fully automatic, and does not allow for any human intervention. The main differences to the other UNLization technologies are the following: 1) SEAN is a multi-document analyser: the input may be not only a single document (as in UNL
editor

and IAN) but a whole collection of documents; 2) SEAN is a word-driven analyser: the unit of analysis is a word (and not a sentence as in UNL
editor

and IAN), to be provided by the user; and 3) SEAN is a shallow analyzer: the analysis targets the surface structure of natural language sentences (and not the deep structure, as in UNL
editor

and IAN). The main consequences of such choices are that the results of SEAN are not appropriate for translation, but for information retrieval and extraction only, because it provides a rather rough and partial analysis of the natural language input. SEAN has been developed by the Library of Alexandria.

NLization

The UNL^{dev} contains only one NLization software, **EUGENE**, the dEep-to-sUrface natural language GENErator. EUGENE is fully automatic. It takes a UNL input and delivers an output in natural language without any human intervention. Similarly to the UNLization tools, it is language-independent and has to be parametrized to the natural language input through a dictionary and a grammar,

provided as separate interpretable files.

Others

NORMA is the UNL normalizer. It is expected to normalize knowledge bases, i.e., sets of statements expressed in the UNL KB format. Normalization, in this case, means suppression of redundancies (relations with the same source and target nodes); suppression of tautologies (relations where the source node is the same as the target node); suppression of contradictions (opposite relations between the same nodes, or the same relations between opposite nodes); generalization (replacing a node by a hyper-node, or a relation by a hyper-relation); specification (replacing a hyper-node by a node or a set of nodes, or a hyper-relation by a relation); merge (replacing two nodes by a single node); and division (replacing one node by two or more nodes). The main goal of the system is to organize and consolidate raw knowledge bases. NORMA has been developed by the Library of Alexandria.

EDGES is the Entity Discovery and Graph Exploration System. It is a UNL visualization tool that displays UNL graphs in hyperbolic mode, and allows for node expansion, collapse and navigation. It is integrated in several front-end applications, as TUT and KEYS, and offers a visual attractive, user-friendly and intuitive way for exploring semantic networks.

FRONT-END APPLICATIONS

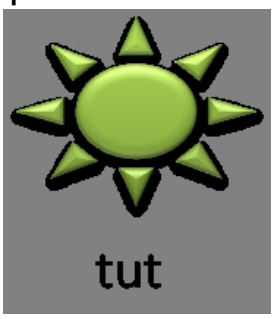
Front-end applications target non-specialists.They are the final destination of the UNL web and constitute the set of end-user systems. The three systems here presented address the most obvious uses of UNL: translation, text processing and information retrieval and extraction. The possibilities of UNL are however unlimited.

LILY is the acronym for Language-to-Interlanguage-to-Language sYstem. It is a machine translation system that uses UNL as a pivot language. It includes IAN as the natural language analysis system, and EUGENE as the natural language generation system. Since IAN is an interactive analysis system, LILY allows for human intervention during the analysis process, but this intervention is rather optional, and the system is prepared to provide results by default, through the highest priorities defined in the dictionary and in the grammar. LILY has to be parametrized for each source and target language, and may include several additional databases from the UNL arium, as the UNL eb, UNL kb

and UNLNL
mb

.

In this sense, it may work either as a knowledge-based MT system or as an example-based MT system. In any case, LILY is always rule-based.



TUT (Text-to-Text through UNL) is a digital library of texts represented in UNL. It comprises links to the integral version of more than 30,000 titles and, whenever available, the UNL version of the text, along with three possible realizations (summarized, simplified and rephrased), in any of the languages available in the UNL System. Its main goal is to UNL-plicate texts. **UNLplication** is the process of generating, through UNL, several different versions from the same source document. These target versions include transformations of language (replication in other languages), of length (text summarization, text extension), of structure (text, matrix, tree, graph) and of social adequacy (text simplification, sociolectalisation). The main goal of the UNLplication process is to reorganize and to reformat the semantic structure of the source document without any explicit commitment to preserving its lexical or syntactic choices, but in a way to extend and enhance its semantic accessibility to a wide range of applications and uses that do not require strict fidelity to the original.

KEYS (Knowledge Extraction sYStem) is a information retrieval and extraction system. It searches for information inside documents represented in UNL, i.e., in semantic hyper-graphs. This allows for retrieval and extraction practices that are language-independent and

semantically-oriented. KEYS includes SEAN, NORMA and EDGES, and is expected to synthesize and normalize the information available on the Web, and to provide summaries extracted out of several different input documents. KEYS has been developed by the Library of Alexandria.