Lexical Issues of UNL

Lexical Issues of UNL:
Universal Networking Language 2012 Panel

Edited by

Ronaldo Martins

CAMBRIDGE
SCHOLARS
PUBLISHING

Lexical Issues of UNL: Universal Networking Language 2012 Panel,
Edited by Ronaldo Martins

This book first published 2013

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-5144-2, ISBN (13): 978-1-4438-5144-2

To Olga Vartzioti (in memoriam)

TABLE OF CONTENTS

List of Abbreviations ix

Preface xi
Tarcísio G. Della Senta

Chapter One 1
Lexical Issues of UNL
Ronaldo Martins

Chapter Two 19
Language Resources: From Infancy to Maturity. Lessons and Next Steps for the UNL Community
Nicoletta Calzolari

Chapter Three 35
Words and Lexical Units
Eric Wehrli

Chapter Four 45
Issues on UWs in the UNL+3 System
Sameh Alansary

Chapter Five 79
On the Possibility of Machine Translation between UNL Dialects: Semantic Units
Igor Boguslavsky

Chapter Six 101
Universal Words and their Relationship to Multilinguality, WordNet and Multiwords
Pushpak Bhattacharyya

Chapter Seven..... 117
Latin Proper Names as UWs
Olga Vartzioti

Contributors..... 135

Index..... 139

LIST OF ABBREVIATIONS

CLARIN	Common Language Resources and Technology Infrastructure
EAGLES	Expert Advisory Group for Language Engineering Standard
ELRA	European Language Resources Association
ENABLER	European National Action for Building Language Engineering Resources
FLaReNet	Fostering Language Resources Network
GCC	GNU Compiler Collection
HLT	Human Language Technologies
HMM	Hidden Markov Model
ISLE	International Standards for Language Engineering
KB	Knowledge Base
LDC	Linguistic Data Consortium
LF	Lexical Function
LMF	Lexical Markup Framework
LR	Language Resources
LT	Language Technology
LW	Language Word
META-NET	Multilingual Europe Technology Alliance
MILE	Multilingual ISLE Lexical Entry
MRD	Machine-Readable Dictionary
MT	Machine Translation
MWE	Multiword Expression
NLP	Natural Language Processing
PANACEA	Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Resources
RDF	Resource Description Framework
UCI	Uniform Concept Identifier
UCL	Uniform Concept Locator
UCN	Uniform Concept Name
UNDL(F)	Universal Networking Digital Language (Foundation)
UNL(C)	Universal Networking Language (Centre)
URI	Uniform Resources Identifier
UW	Universal Word

PREFACE

This book inaugurates a series of discussions on what is permanent in the original thinking of Universal Networking Language (UNL), and the changes introduced during its development. The purpose is to highlight the UNL's fundamental principles that remain as firm as when they were formulated several years ago, while their materialization evolves over time, following the advances in Linguistics, Knowledge Engineering and Information Sciences.

Remaining unchanged in UNL are:

- The idea of an artificial language that is able to describe the universe as any other human language;
- The idea of a language that, though artificial, is made of lexical, grammatical and semantic components as any natural language;
- The idea of a language that can represent information and knowledge independently of natural languages;
- The idea that it is a language for machines, which enables human-machine interaction in an intelligent partnership.

For more than a decade, eminent linguists, IT developers and NLP scholars have worked together on the materialization of the "idea" of UNL. To begin with, they adopted set specifications on the formalism of the UNL that were followed by all. As their work progressed, they gradually realized the need for adjusting some of the initial specifications and for introducing new ones.

These specifications concern three basic components of the UNL linguistic structure: The "Universal Words" (UW) which constitute the vocabulary of the UNL; the "Relations" that describe semantic functions between two UWs; and "Attributes" that describe circumstances under which UWs and Attributes are used.

Of course, changes in the specifications for these three components have often been proposed but not uniformly adopted by all members of the UNL Community. In this context, the UNDL Foundation launched a series of panels to discuss issues and problems that emerge while developing UNL. Eminent scholars working on UNL and researchers in the domain of

Linguistics and Natural Language Processing are invited to analyze and advise on changes that would improve the UNL formalism.

The I UNL Panel was dedicated to discuss the concept and practice of "universal word" (UW). It was held in Mumbai in December 2012, at the same time as COLING. Six eminent scholars were invited to present papers and discuss them in front of other experts, including some of those involved in developing the UNL. They were asked to elaborate five basic questions ranging from basic specifications of UWs in UNL formalism to the practical use in specific linguistic phenomena, such as proper names and multiword expressions.

This book reflects the papers presented and the debates among the panellists and participants that followed. It presents a broad spectrum of current issues and technical developments in NLP related in particular to the UNL Interlingua approach, and proposes some practical answers to the five questions submitted to the I UNL Panel.

The book is composed of seven chapters. In the first, Professor Ronaldo Martins, Language Resources Manager of the UNDL Foundation and main editor of this book, sets the context of the whole discussion. He explains what the mission of the UNL Program is and what UNL itself is committed to. He then proposes a clarification on how to understand the submitted questions by introducing what is meant by Universal Words, their types, their properties and their structure. Each panellist subsequently confronts their own opinion on the understanding of Universal Words by answering directly or indirectly, the five questions asked during the meeting in Mumbai.

In the second chapter, Professor Nicoletta Calzolari Zamorani, Research Associate and former Director of CNR-ILC, Pisa, Italy, approaches the questions by giving insight on the general framework surrounding Language Resources (LRs) today. Rather than answering the specific UNL questions directly, she emphasizes the fact that in order to define a language structure such as UNL it is pree-eminent to gather results of analyses from different communities and to define a coherent strategy.

The third chapter highlights the issues of "lexical units" in machine translation by focusing on multiword expressions, and collocations in particular. Professor Eric Wehrli, from the University of Geneva, Switzerland, attempts to determine what should be the proper level of representation for lexical transfer in machine translation. Going through multiple examples of multiword expressions and collocations he offers the background for approaching the UNL-based questions.

Professor Sameh Alansary, head of the Arabic UNL Centre at the Library of Alexandria, Egypt, on the other hand, answers the five questions directly in Chapter Four. His categorical analysis of the implications of the proposed questions reflects the preciseness with which Universal Words must be handled in an interlingual context. After his comprehensive, but moreover profound study, Alansary provides us with straight answers to the five proposed questions.

In Chapter Five, Professor Igor Boguslavsky, from the Russian Academy of Sciences, Moscow and the Universidad Politécnica de Madrid, handles the five questions under a different light. He compares the different approaches to the UNL development in the last ten years or so by three institutions: the UNDL Foundation, the UNL Centre and the U++ Consortium. Each of these has introduced some changes in the specifications for encoding and decoding the UNL. Boguslavsky answers the five questions by highlighting these differences, positioning himself on the side of the U++ Consortium approach.

In Chapter Six, Professor Pushpak Bhattacharyya, from the Indian Institute of Technology Bombay (IITB) Mumbai IIT, India, points to the fact that the UW dictionary created by the UNDL Foundation is based on the English WordNet. To him, this poses some special issues regarding multilinguality and multiwords in the construction of the UW dictionary itself. According to him, this may hinder the UNL's ambition of being a universal repository of lexical knowledge.

Professor Olga Vartzioti, from the Department of Philology at the University of Patras, Greece, gives us a more theoretical approach of the case in Chapter Seven. She bases her analysis on proper names and in particular on Latin proper names, giving us a philologist's and historical linguist's view on the matter.

In summary, the five questions prompted the contributions of scholars from different language backgrounds and approaches from various schools of thought. The book reflects the variety of their thinking on the UW concept in general, as well as their suggestions for tackling some of the problems it poses in practice. The UNDL Foundation is grateful for their contributions and welcomes them all in future panels.

Geneva, July 4th 2013
Tarcísio G. Della Senta
UNDL Foundation, President

The UNDL Foundation is grateful to the Arab Fund for Economic and Social Development of Kuwait for the support provided to the I UNL Panel and this publication.

CHAPTER ONE

LEXICAL ISSUES OF UNL

RONALDO MARTINS

Introduction

The Universal Networking Language (UNL) is a knowledge representation language that has been used in several different fields of Natural Language Processing, such as Machine Translation, multilingual document generation, summarization, information retrieval and extraction, sentiment analysis and semantic reasoning. It was proposed by the Institute of Advanced Studies of the United Nations University, in Tokyo, Japan in 1996, and has been developed by the UNDL Foundation in Geneva, Switzerland under a mandate of the United Nations since 2000.

Originally designed more than fifteen years ago, the UNL has not escaped the action of time and has not yet incorporated several recent advances in the domain of Natural Language Processing. In order to prepare the ground for the necessary updates to the present specifications, the UNDL Foundation set up the UNL Panel initiative, whose main purpose is to collect the opinion of specialists, from inside and outside the UNL Community, about the desirable structure, nature and role of the elements of the UNL. In order to organize the discussion, the UNDL Foundation divided the subjects into three chapters to be addressed in three different meetings:

- Universal Words (the set, notation and properties of UWs);
- Relations and attributes (the set, notation and properties of relations and attributes);
- UNL document structure (format, encoding, schema and validation).

The first meeting took place in Mumbai on December 15 2012 as a collocated event to COLING 2012, and was dedicated to Universal Words (UWs). During this meeting, six specialists were requested to address five

general questions about the repertoire, the structure and the role of UWs inside the UNL framework:

- What is to be considered a "Universal Word" (UW)?
- Which named entities should be introduced in the dictionary of UWs, if any?
- UWs must correspond to roots, to stems or to word forms?
- Antonyms should be represented as a single UW or as different UWs?
- When a multiword expression must be represented as a UW?

After a general introduction by the UNDL Foundation, each panellist delivered a thirty minute oral presentation in which they discussed which answers would be more appropriate and feasible in each case, considering the state of the art of the theory and technology on Natural Language Processing. They also suggested some general procedures to be adopted in similar cases, either confirming, and sometimes refusing, our current practices, which were the object of lively discussions. The present publication is the written version of these presentations, and is intended to subsidise an in-depth revision of the current specifications.

This first chapter is dedicated to presenting the set of assumptions and existing specifications of the UNL presented by the UNDL Foundation in the introductory session as the common ground for the debate. It is divided into three sections. The first section demonstrates the commitments of the UNL, the keystones of the language, which were expected to be taken for granted by all the participants. The second section presents the current position concerning the Universal Words, the main goal of the I UNL Panel, which the panellists were free to amend and criticize. The third and last section provides the detailed presentation of each question with the corresponding theoretical and practical issues.

The UNL

The mission of the UNL Program is to construct an artificial language – the UNL – that can be used to process information across language barriers. The major commitments of the UNL are the following:

The UNL must represent information

The UNL is an artificial language designed to represent information. In this sense, the UNL is primarily a knowledge representation language. The

most important corollary of this first commitment is that UNL is not intended to describe or represent natural languages. It is used to describe and represent the information conveyed by natural languages. The goal of UNL is to represent "what was meant" and not "what was said" or "how it was said.» Accordingly, UNL provides an interpretation rather than a translation of a given utterance. The UNL version of an existing document is not committed to preserving the lexical and the structural choices of the original, but must represent, in a non-ambiguous format, one of its possible meanings, preferably the most conventional one.

The UNL must be a language for computers

The UNL is an artificial language shaped to represent information in a machine-tractable format. Like other formal systems, it seeks to provide the infrastructure for computers to handle what is meant by natural languages. Different from other auxiliary languages (such as Esperanto, Interlingua, Volapük, Ido and others), the UNL is not intended to be a human language. We do not expect people to speak UNL or to communicate in UNL. The UNL, like HTML, SGML, XML and other markup languages, is a language for machines.

The UNL must be self-sufficient

In the UNL approach there are two basic movements: UNLization and NLization. UNLization is the process of representing the information conveyed by natural languages into UNL; NLization, conversely, is the process of generating a natural language document out of UNL. In order to be fully tractable by machines, the UNL must be self-sufficient, i.e. should be as semantically complete and saturated as possible. The UNL representation must not depend on any implicit knowledge and should explicitly codify all information. This means that the UNLization should be completely independent from the NLization, and vice-versa: the UNLization should not take into consideration the target language or format of any future NLization and the NLization should not need any information about the source language or previous structure of any UNL document.

The UNL must be general-purpose

At first glance, the UNL seems to be a pivot-language to which the source texts are converted before being translated into the target languages.

It can, in fact, be used for such a purpose, but its primary objective is to serve as an infrastructure for handling information. In addition to translation, the UNL is expected to be used in several other different tasks, such as text mining, multilingual document generation, summarization, text simplification, information retrieval and extraction, sentiment analysis etc. Indeed, in UNL-based systems there is no need for the source language to be different from the target language: An English text may be represented in UNL in order to be generated, once again, in English, as a summarized, simplified, localized or simply rephrased version of the original.

The UNL must be independent from any particular natural language

The UNL is expected to be the language of the United Nations and, therefore, must not be circumscribed to any existing natural language in particular under the risk of being rejected by the state members of the General Assembly.

These five commitments materialized in a formal system first presented in 1996[1] and amended several times since then.[2] The most recent specification is available at www.unlweb.net/wiki/Specs. The cornerstones of the current implementation are the following:

The UNL represents information through semantic networks

The UNL assumes that information can be represented as a graph structure (a "semantic network") composed of nodes (concepts) and arcs between nodes (semantic relations between concepts). This idea is not new. Semantic networks have been used in knowledge representation at least since Charles S. Peirce[3] and as an Interlingua for machine translation since 1956.[4]

---

[1] UNL. (1996). Universal Networking Language: an electronic language for communication, understanding and collaboration. Tokyo: UNL Center.
[2] Version 1.0 (April, 1998), Version 1.5 (May, 1998), Version 2.0 (July, 1999), Version 3.0 (November, 2001), Version 3.1 (May, 2002), Version 3.2 (July, 2003), Version 3.3 (December, 2004) – All these versions may be downloaded at http://www.unlweb.net/wiki/index.php/Specs.
[3] Peirce (1909).
[4] Richens (1987).

The UNL includes a universal lexicon

Languages are normally defined as the combination of a lexicon (i.e. a set of words) and a grammar (i.e. a set of rules for combining words). As a "universal language" the UNL includes a "universal lexicon" and a "universal grammar". The universal lexicon is derived from the assumption that words from different languages may be co-indexed through common semantic entities (the "Universal Words" or "UWs"), which would stand as linguistic counterparts for the notion of "universal concept". The idea of universality in language, and of a universal dictionary, is as old as the language studies.[5] However, except for extremely constrained domains (such as Chemistry), none of these attempts really succeeded up to now, including the UNL, whose set of UWs has been subject to continuous changes, both in the form (i.e. in the way of expressing UWs) and content (the actual list of UWs).

The UNL includes a set of semantic binary relations

The "universal grammar" of UNL consists of a set of relations and attributes. The set of relations comprises semantic binary relations that link UWs in order to form the UNL graph. This set of relations (currently between forty and fifty) has also been undergoing several changes since 1996 and this is the main difference between the several versions of the UNL specifications.[6] The idea of semantic binary relations has been proposed in numerous linguistic approaches, the most famous being the "structural syntax" developed by Lucien Tesnière in the 1930s, and the "semantic case" presented by Charles Fillmore in 1968. Although the lists of semantic relations vary considerably, they normally share some basic common elements: "agent", "patient", "instrument", "place", "time" etc.

The UNL includes a set of semantic attributes

The "universal grammar" of UNL also contains attributes which are used to specify the use of UWs and to assign pragmatic information to UNL statements. The current set of attributes represents mostly information about aspect (@habitual, @persistent, @progressive, etc.),

[5] Umberto Eco in *La Recherche de la langue parfaite dans la culture européenne* (1993) provides a comprehensive account of concrete and documented attempts to build a “perfect language”, most of them based on the idea of semantic primitives or a universal dictionary, dating at least from Ramon Llul (thirteenth century).

[6] These versions are available at http://www.unlweb.net/wiki/index.php/Specs.

degree (@minus, @plus, etc.), emotions (@anger, @fear, etc.), gender (@male, @female), modality (@hypothesis, @obligation, @necessity, etc.), position (@proximal, @distal), quantification (@paucal, @multal, @plural), register (@taboo, @colloquial, etc.), solidarity (@polite, @intimate, etc.) and time (@past, @present, @future).

The UNL is a markup language

The UNL is used for annotating a natural language document in a language-independent (semantically-based) way. This is done through the UNL document structure, which can be illustrated by the example below:

```
[S:1]
{org:eng}
Peter kissed Mary
{/org}
{unl}
agt(kiss(agt>thing,obj>thing).@past,Peter(iof>person))
obj(kiss(agt>thing,obj>thing).@past,Mary(iof>person))
{/unl}
[/S]
```

In the example above:

[S] and [/S] indicate the beginning and the end of the sentence, respectively

{org} and {/org} indicate the beginning and the end of the original sentence, respectively

{unl} and {/unl} indicate the beginning and the end of the UNL graph, respectively

"agt" (agent) and "obj" (object) are semantic relations of UNL

"kiss(agt>thing,obj>thing)","Peter(iof>person)"and "Mary(iof>person)" are UWs

@past is an attribute

As indicated above, the main features of UNL are not actually unique, and have long been part of research and development in the field of Computational Linguistics and Artificial Intelligence. The originality of the current approach rather concerns the combination of these features in a single representation, i.e. that the UNL semantic network must be composed with a specific universal lexicon, with a specific set of semantic

relations and with a specific set of semantic attributes, represented in a given format. This is the keystone of the UNL and its main distinctive feature in comparison to other semantic networks and formal systems. Provided that the concepts deployed in the UNL approach (semantic networks, universal dictionary, relations, attributes) have been used in the field of Natural Language Processing for a long time we understand that the novelty of UNL does not concern "what" it does, but "how" it does what it does i.e. "with which resources" it does it.

Universal Words

The basic assumption of the UNL approach is that the information conveyed by natural languages can be formally represented by a semantic network made up of three different types of discrete semantic units: Universal Words (UWs), relations and attributes. The UWs are the nodes of the network and are interlinked by relations and specified by attributes. This three-layered representation is the cornerstone of the UNL. In this book, we will concentrate on the UWs.

As the name indicates, UWs are expected to be "universal". This does not mean that they represent a sort of common lexical denominator to all languages or a semantic primitive. The concept of "universality," in UNL, must be understood in the sense of "capable of being used and understood by all," and UWs depict concepts that may range from global to local and even temporary. UWs may represent concepts that are believed to be lexicalized in most languages (such as "cause to die"); concepts that are lexicalized only in a few languages (such as "to execute someone by suffocation"); concepts that are lexicalized in one single language (such as "a person who is ready to forgive any transgression a first time and then to tolerate it for a second time, but never for a third time"); and concepts that are not lexicalized in any language (such as "women that normally wear red hats and white shoes in big theatres"). The universality of a UW does not come from the type of concept that it represents, but from the way it does it. The UW provides a method for processing the concept, so that any natural language would be able to deal with it, either as a single node, if lexicalized, or as a hyper-node (i.e. a sub-graph).

Types of UWs

UWs can be permanent or temporary. Permanent UWs are included in the UNL Dictionary and correspond to concepts that are lexicalized in at

least one language (i.e. which are conceived as single lexical items and are therefore included in natural language dictionaries). Temporary UWs are words that represent concepts or entities that are still in the process of lexicalization (e.g. "googlers" or "twittered"), are too specific to be included in the UNL Dictionary ("Universal Networking Digital Language Foundation", "Léon Werth"), or that are not translatable ("3.14159", "H_2O", "www.undlfoundation.org").

Permanent UWs can be simple, compound or complex. A simple UW is an isolated node in the UNL graph. It is used when the UW represents a concept that is not compositional, i.e. that cannot be fully broken into constituent concepts, such as "big(icl>size)" (= a value of size) or "stamp(icl>token)" (= a type of token). A compound UW is an isolated node combined with attributes. It is used when the concept can be decomposed into an existing simple UW and an attribute, such as "big(icl>size).@more" (= bigger). A complex UW is a hyper-node, i.e. a sub-graph inside the UNL graph. It is used when the concept can be reduced to the combination of existing UWs, relations and attributes, such as "obj(to affix(icl>to attach),stamp(icl>token))" (= to stamp, i.e. to affix a stamp).

Properties of UWs

In the UNL framework, the following properties apply to UWs:

Sense

UWs represent sense and not reference. UWs are related to the intension (sense, meaning and connotation) rather than to the extension (reference, denotation) of linguistic expressions. The expressions "morning star" and "evening star", which are said to have the same reference (the planet Venus), must be necessarily represented by different UWs because they convey different modes of presentation of the same object, i.e. they have different senses: "the last star to disappear in the morning" and "the first star to appear in the evening", respectively.

Productivity

UWs must correspond to, and only to, contents conveyed by natural language open lexical categories (nouns, verbs, adjectives and adverbs). Any other semantic content (such as the ones conveyed by articles, prepositions, conjunctions etc.) should be represented as attributes or relations. This criterion is not language-biased: If a given semantic value

proves to be conveyed, in any language, by a closed class, it should not be represented as a UW, regardless of its realization in other languages.

Compositionality

Simple UWs must correspond to and only to contents expressed by non-compositional lexical items, i.e. words and multiword expressions that cannot be reduced to the combination of existing UWs, attributes and relations. Compound and complex UWs must be used when the content can be fully determined by the meanings of constituent expressions and the rules used to combine them.

Comprehensiveness

UWs are "universal" in the sense that they constitute the lexicon of a "universal language", i.e. that they convey ideas that can be expressed in each and every language. They are not universal in the sense that they are lexicalized in all languages. In that sense, UWs are not to be considered semantic primitives, nor should they represent only common concepts. The repertoire of UWs is supposed to be as comprehensive as the set of different individual concepts depicted by different cultures, no matter how specific they are. Furthermore, the lexicon of UNL constitutes an open set, subject to permanent increase with new UWs, as UNL is supposed to incessantly incorporate new cultures and cultural changes.

Universality

Permanent UWs may represent concepts with different degrees of universality and are stored accordingly in three nested lexical databases, which are subdivisions of the UNL Dictionary:

- The UNL Core Dictionary contains only permanent simple UWs that represent concepts that are (presumably) lexicalized in all languages
- The UNL Abridged Dictionary contains all permanent UWs (simple, compound or complex) that represent concepts that are lexicalized in at least two different language families
- The UNL Unabridged Dictionary contains all permanent UWs (simple, compound or complex) that represent concepts that are lexicalized in at least one language

Non-Ambiguity and Non-Redundancy

A given sense may not be represented by more than one UW, and one UW may not have more than one sense. There is no homonymy, synonymy or polysemy in UNL.

Arbitrariness

Simple UWs are names (and not definitions) for senses, and the simple UW does not bring much (or any) information about its sense. It is just a label. Any information concerning the sense is expected to be provided by the three different lexical databases available inside the UNL framework: the UNL Dictionary, the UNL Knowledge Base and the UNL Memory.

Structure of UWs

Temporary UWs (such as "www.undlfoundation.org", H_2O, +41 (0) 22 879 80 90, etc.) are always represented between double quotes, and observe the source language spelling practices (concerning, for instance, capitalization). For the time being, they are also expected to be transliterated in Roman script.

Permanent UWs are represented through Uniform Concept Identifiers (UCI). The UCI is a particular type of Uniform Resource Identifier (URI) used to refer to UWs. In the UNL framework, UCIs are represented either as UCL (Uniform Concept Locator) or UCN (Uniform Concept Name).

Uniform Concept Locators (UCL), like URLs, provide a method for finding the concept in the UNL Knowledge Base. They are represented as:

ucl://<AUTHORITY>/<ID>

Where:

- ucl is the scheme name for uniform concept locators
- <AUTHORITY> is the authority (knowledge base) responsible for the concept (unlkb.unlweb.net, by default)
- <ID> is the index of the concept in the knowledge base

For instance, the concept "a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs", which is lexicalized in English through the noun "table", may be located through:

ucl://unlkb.unlweb.net/104379964

This address brings all the information concerning the concept, i.e. its definition in UNL, which may be used by the languages where this concept is not lexicalized.

Uniform Concept Names (UCN) use the ucn scheme and, as URNs, do not imply availability of the identified resource. They are represented as:

ucn:<LID>:<NSS>
Where:

- ucn is the scheme name for uniform concept names
- <LID> is namespace identifier, which corresponds to the three-character ISO 639-2 code for languages
- <NSS> is the namespace-specific string

For instance, the same concept depicted above (corresponding to "table" in English) may be associated to several different UCNs:

ucn:eng:table(icl>furniture)
ucn:fra:table(icl>mobilier)
ucn:esp:mesa(icl>mobiliario)
ucn:deu:Tisch(icl>Möbel)
ucn:rus:стол(icl>мебель)

UCN's are unique in each language and the namespace-specific string is normally split into two different parts: a root and a suffix, as exemplified above. The root ("table", "mesa", "Tisch", "стол") can be a word or a multiword expression. The suffix ("icl>furniture", "icl>mobilier", "icl>mobiliario", "icl>Möbel", "icl>мебель") is always introduced by a UNL relation (e.g.: "icl" = is a kind of) and is used to disambiguate the root.

In the UNL Document Structure, UCIs are always abbreviated to the last part, because the scheme, the authority and the namespace may be inferred from the document header. For instance:

- 104379964 instead of ucl://unlkb.unlweb.net/104379964
- table(icl>furniture) instead of ucn:eng:table(icl>furniture)

Type	Concept (in English)	Lexicalization (in English)	UCL	UCN (in English)
Simple UW	above average	big	301382086	big(icl>size)
Compound UW	comparative of above average	bigger	301382086.@more	big(icl>size).@more
Complex UW	to affix a stamp	to stamp	obj(201356370,106796119)	obj(to affix(icl>to attach), stamp(icl>token))
Temporary UW	UNDL Foundation	UNDL Foundation	"UNDL Foundation"	

Table 1-1 - Examples of UWs

Open lexical issues in UNL

During the I UNL Panel, the UNDL Foundation listened to six specialists regarding five topics of lexical semantics:

- What is to be considered a "Universal Word"?
- Which named entities should be introduced in the dictionary of UWs, if any?
- UWs must correspond to roots, to stems or to word forms?
- Antonyms should be represented as a single UW or as different UWs?
- When a multiword expression must be represented as a UW?

In order to normalize the discussion, these topics were instantiated in the five questions below. They illustrate practical issues concerning UWs and have been received several different possible answers. The main goal of I UNL Panel was to discuss which answers would be more appropriate and feasible, considering the nature and role of the UNL, and the state of the art of the theory and technology on Natural Language Processing. Participants were expected to use those particular cases as starting points for their presentations and to suggest some general procedures to be adopted in similar cases, which could either confirm or deny our current practices. The five questions were the following:

How many UWs should be recognized in the sentence below?

"Charles Dickens is generally regarded as the most important English novelist of the Victorian period"

The basic assumption of the UNL approach is that the information conveyed by natural languages can be formally and usefully represented through semantic networks composed of three different types of discrete semantic entities: UWs, relations and attributes. UWs are nodes in the UNL graph; relations are the arcs between nodes; and attributes are specifiers that restrict the extension of nodes. This three-layered representation poses several problems to the UNLization as the distinction between these three entities is not always clear. Consider, for instance, the sentence given above. How many UWs (either permanent or temporary) should be recognized in this sentence?

- "Victorian period" should be represented as single UW ("Victorian period") or as two different UWs ("Victorian" and "period")?
- The verb "to be" should be represented as a UW or as a relation between "Charles Dickens" and "the most important English novelist of the Victorian period"? (Consider also the options "was" and "has been" in the same context).
- The preposition "of" should be represented as a UW or as a relation between "the most important novelist" and "the Victorian period"? (Consider also the options "since", "from ... on", "in" or "during" instead of "of").
- "generally regarded as" should be represented by UWs ("generally", "regarded", "as", for instance) or as an attribute (a downtoner, which lowers the truth effect of the declaration) to be assigned to the whole proposition "Charles Dickens is the most important English novelist of the Victorian period"?
- The adverb "most" should be represented as a UW or as a superlative marker (to be represented as an attribute to be assigned to the adjective "important"?) (Consider also "greatest English novelist" instead of "most important English novelist").

"Charles Dickens" should be represented as a permanent UW or as a temporary UW?

The UNL Dictionary contains only permanent UWs. Untranslatable expressions, though transliterable, are not included in the dictionary, but may be used in the UNL graphs as temporary UWs. This is the obvious case for URLs, e-mail addresses, phone numbers, formulae, etc. However, there are cases in which these criteria are still under dispute: proper names (of people, of places, of brands, etc.) for instance. When should they be considered permanent UWs (and included in the UNL Dictionary) and when should they not? Consider, for instance, the case of "Charles Dickens". Should it be defined as a permanent UW and included in the UNL Dictionary? Or should it be treated as a temporary UW? Consider also the cases of "Charles J Dickens" (an American citizen born on June 17, 1949 and died on October 21, 2004); the "Charles Dickens Museum" located in London; the bar and restaurant "Charles Dickens" located in Southwark; the "Charles Dickens School" located in Kent; and other entities named "Charles Dickens". Consider the size (and the maintenance) of the UNL Dictionary, in case you suggest to treat them all as permanent

UWs; or, otherwise, consider how to handle concepts that have not been included in the UNL Dictionary.

"hunger" (= "a physiological need for food"), "hungry" (= "feeling hunger"), "hungrily" (= "in the manner of someone who is very hungry") and "hunger" (= "to cause to experience hunger") should be represented as simple, compound or complex UWs?

In the current framework, UWs can be simple, compound or complex. A simple UW is represented as a node in the UNL graph. A compound UW is represented as a node with attribute(s). A complex UW is represented as a sub-graph, i.e. as a set of interlinked nodes. This offers different possibilities in representing the concepts above. For instance:

Simplified UW candidates for "hunger", "hungry", "hungrily" and "to hunger"			
Lexical Item (English)	**Simple UW**	**Compound UW**	**Complex UW**
hunger	hunger	hungry.@ness	a physiological need for food
hungry	hungry	hunger.@full_of	feeling hunger
hungrily	hungrily	hunger.@full_of.@manner hungry.@manner	in the manner of someone who is very hungry
hunger	hunger	hunger.@full_of.@make hungry.@make	to cause to experience hunger

Table 1-2 – Simplified UW candidates

Which is the best way to represent these concepts? Consider the fact that some of these concepts are not lexicalized in all languages (for

instance, the adjective "hungry" is not very frequently found in German and French: "I am hungry" is normally translated as "Ich habe Hunger" or "J'ai faim", respectively). Consider also the actual importance of part-of-speech for lexical semantics. Consider, at last, the actual "compositionality" of these concepts.[7]

Antonyms such as "mortal" and "immortal", "hot" and "cold", and "son" and "father" should be represented as a single UW (and the corresponding attributes) or as different UWs?

The UNL is expected to be non-redundant: synonyms (such as "hunger" and "hungriness") and paraphrases (such as "Mary killed Peter" and "Peter was killed by Mary") are expected to be represented in the same way. What should we do with antonyms? Should we have a non-marked UW (such as "mortal", "hot" and "son") and generate their antonyms as compound UWs (such as "mortal.@not", "hot.@not" and "son.@converse") to avoid vocabulary multiplication and to cover languages with lexical gaps (unpaired words)? Or should we represent all of them as simple UWs ("mortal", "immortal", "hot", "cold", "son", "father") because they could not be fully reduced to the combination of a

[7] It is important to stress that these differences do not pose any practical restrictions to the UNL representation. For instance, the English noun phrase "hungry boy" could be represented in UNL as:

- mod(boy, hungry) ("feeling hunger" as a Simple UW)
- mod(boy, hunger.@full_of) ("feeling hunger" as a Compound UW)
- mod(boy, :01)obj:01(to feel,hunger) ("feeling hunger" as a Complex UW)

In the same way, these differences do not pose any restrictions to the resources (dictionaries and grammars). For instance, the French dictionary could bring:
[affamé]{} "hungry" (LEX=J,POS=ADJ,GEN=MCL,NUM=SNG)<fra,0,0>; ("delighting the senses" as a Simple UW)
[affamé]{} "hunger.@full_of" (LEX=J,POS=ADJ,GEN=MCL,NUM=SNG)<fra,0,0>; ("delighting the senses" as a Compound UW)

- [affamé]{} "obj(to feel, hunger)" (LEX=J,POS=ADJ,GEN=MCL,NUM=SNG)<fra,0,0>; ("delighting the senses"as a Complex UW)

But these differences do pose semantic consequences: A simple UW represents a concept seen as a single unit, whereas compound and complex UWs are strictly compositional, i.e. the meaning of the UW is entirely derived from its components. Furthermore, translating "I am hungry" by "Je suis affamé", although possible, is not convenient in French.

simple UW and an attribute? Consider the case of absolute opposites (such as "mortal" x "immortal" which could be opposed by an attribute such as @not), of gradable opposites (such as "hot" and "cold" which would also require intensifiers such as hot.@extra, hot.@plus, hot, hot.@minus, hot.@not, hot.@not.@minus, hot.@not.@plus and hot.@not.@extra), and of relational opposites (such as the converse "son" and "father" that would require a special attribute - @converse, for instance - to inform that if x is the son of y, y is the father of x).

"Farbfernsehgerät" ("color television set", in German) should be represented as a simple or complex UW?

According to current standards, every concept lexicalized in at least one language must be defined as a permanent UW and included in the UNL Dictionary. The concept of "lexicalization" is, however, highly controversial, and seems to vary considerably between different languages, and even between different lexicographical approaches for the same language. This has been true especially for multiword expressions, i.e. lexemes containing more than one stem, which are recognized as single entries in some dictionaries, and simply ignored by others. For the time being, we have been avoiding this discussion by assuming that if a word was included (either as an entry or a sub-entry) in any knowledgeable dictionary, it should be considered "lexicalized" and, therefore, defined as a permanent UW. But this procedure seems to be exaggeratedly language-dependent. "Farbfernsehgerät" for instance, is considered to be lexicalized in German because it can be found in German dictionaries as one single entry; the English equivalent, "colour television set" seems however not to be lexicalized in English yet because it could not be found in the major English dictionaries. Should we represent this concept as a simple (non-compositional) UW (as in German), or as a complex (compositional) UW (as in English)? Consider the fact that "Farbfernsehgerät" is formed by "Farbe", "Fernsehen" and "Gerät", i.e. that the compound is not simply the concatenation of three words, but has undergone spelling changes (in addition to semantic changes, if any). Consider also the case of compounds such as "baby-talk" (tatpuruṣa or endocentric, i.e. "baby" is a special kind of "talk"), "bittersweet" (dvandva or copulative, i.e. "bitter" and "sweet") and "skinhead" (bahuvrihi or exocentric, i.e. non-compositional). Consider, finally, the case of idioms, such as "all ears", "closed book" and "cold feet".