Le Petit Prince in UNL

Ronaldo Martins
UNDL Foundation
48, Route de Chancy, CH-1213, Petit-Lancy, Geneva, Switzerland
r.martins@undlfoundation.org

Abstract

The present paper addresses the process and the results of the interpretation of the integral text of "Le Petit Prince" (Little Prince), the famous novel by Antoine de Saint-Exupéry, from French into UNL. The original text comprised 1,684 interpretation units (15,513 words), which were sorted according to their similarity, from the shortest to the longest ones, and which were then projected into a UNL graph structure, composed of semantic directed binary relations linking nodes associated to the synsets of the corresponding original lexical items. The whole UNL-ization process was carried-out manually and the results have been used as the main resource in a natural language generation project involving already 27 languages.
.

Keywords: UNL, semantic annotation, knowledge representation

1. Introduction

The Universal Networking Language, or UNL (Uchida, Zhu & Della Senta, 1999, 1996) 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 semantic reasoning. It was originally proposed by the Institute of Advanced Studies of the United Nations University, in Tokyo, Japan, and has been currently promoted by the UNDL Foundation, in Geneva, Switzerland, under a mandate of the United Nations.

In the UNL approach, the information conveyed by natural language is represented, sentence by sentence, as a directed graph in which nodes represent concepts, and edges represent binary semantic relations between concepts. The nodes are called Universal Words (or simply UWs), and may be further specified by a predefined set of attributes which cover the information normally represented by closed-class categories (such as tense, aspect, number and gender, for instance). The set of binary relations is also predefined in the UNL Specifications and consists of 46 semantic cases (such as agent, object, instrument, etc.).

The UNL-ization process consists in mapping the information that is verbally elicited in the surface structure of written texts into UNL. The resulting graph, although preserves the sentence boundaries defined in the source document, is not committed to replicate the lexical and the syntactic choices of the original, but focuses in representing, in a non-ambiguous format, one of its possible readings, preferably the most conventional one. In this sense, the UNL representation is said to be an "interpretation" rather than a "translation" of a given text.

The UNL-ized version of a text may be used in several different applications, such as machine translation (if the target language is different from the source language); simplification (if the target language is a simplified version of the source language); summarization (if the graph is summarized prior to natural language generation); knowledge extraction (if the graph is normalized by reference to its nodes); etc.

The present paper addresses the process and the results of the UNL-ization of "Le Petit Prince" (Little Prince), the famous novel by Antoine de Saint-Exupéry, first published in French in 1943. "Le Petit Prince" is one of the best-selling books ever (more than 80 million copies), and has been translated to more than 180 languages, providing thus the possibility of contrasting and evaluating a wide range of UNL-based translations. Additionally, the text offers the chance of experimenting UNL in three situations that have not been explored so often: French original, narrative and literature. Our main goal was to regenerate the text in at least three different directions: replication, summarization and simplification, in as many languages as possible.

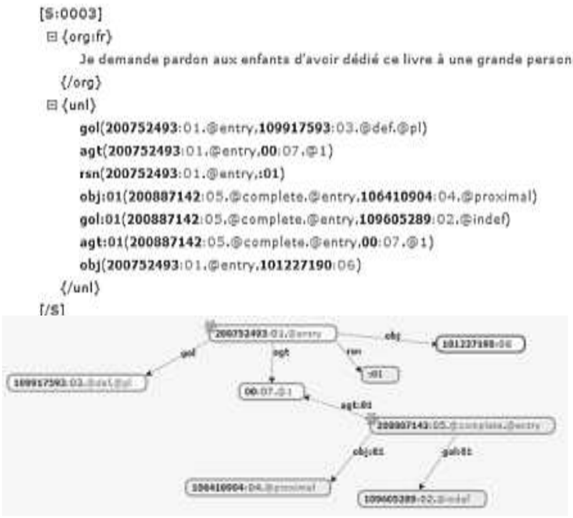
2. The UNL-ization of Le Petit Prince

The integral version of Le Petit Prince, which has been released under public domain in Canada, was obtained from wikilivres.info/wiki/Le_Petit_Prince. The whole text comprises 15,513 word forms (tokens) and 1,684 sentences. The UNL-ization of the text was carried out in a fully-manual way through the UNL Editor, a graph-based authoring tool developed by the UNDL Foundation and available at the UNLdev (www.unlweb.net).

The sentences were divided into two main different groups: a) the training corpus, which comprises the first 53 sentences of the book (dedication and first chapter), including the title; and b) the application corpus, which comprises the remaining 1,548 sentences. The training corpus was addressed collectively by a group of four human UNL-izers in order to synchronize and normalize the UNL-ization strategies. The application corpus was organized according to the similarity of sentences (and not to the order of appearance) and was addressed from December 2009 to February 2010 according to the guidelines resulting from the training exercise, which were collected at www.unlweb.net/wiki/index.php/UNLization_Guidelines.

As the main goal of the UNL-ization process was to represent the text in a language-independent and non-ambiguous format, the source document was fully-normalized: all semantic valences were saturated, including anaphora, ellipses, presuppositions and implicatures. Pronouns and pro-forms, for instance, were replaced by their antecedents, and were represented in UNL only in case of exophoric reference (indefinite pronouns, interrogative pronouns and personal pronouns that were not co-indexed to any existing antecedent). The normalized sentences were represented as graphs in which open lexical categories (nouns, verbs, adjectives and adverbs) were represented as nodes (UWs) indexed to synsets extracted from the English WordNet (version 3.1). Bound morphemes and closed classes – i.e., affixes (gender, number, tense, aspect, mood, voice, etc), determiners (articles and demonstratives), adpositions (prepositions, postpositions and circumpositions), conjunctions, auxiliary and quasi-auxiliary verbs (auxiliaries, modals, coverbs, preverbs) and degree adverbs (specifiers) – were represented as attributes to the nodes. Attributes were also used to deal with figurative language (mainly metaphor) and non-verbal elements of communication (such as politeness, schemes and speech acts). The nodes were interlinked by the directed binary relations extracted from the UNL 2005 Specs (available at www.undl.org/unlsys/unl/unl2005/).

The whole corpus was later uploaded to the UNLarium (www.unlweb.net/unlarium), a web-based database management system where anyone is able to browse and to export it. A sample of the UNL-ization result is presented below in two different formats (table and graph):



```
[S:0003]
 ⊟ {org:fr}
     Je demande pardon aux enfants d'avoir dédié ce livre à une grande personne.
   {/org}
 ⊟ {unl}
     gol(200752493:01.@entry,109917593:03.@def.@pl)
     agt(200752493:01.@entry,00:07.@1)
     rsn(200752493:01.@entry,:01)
     obj:01(200887142:05.@complete.@entry,106410904:04.@proximal)
     gol:01(200887142:05.@complete.@entry,109605289:02.@indef)
     agt:01(200887142:05.@complete.@entry,00:07.@1)
     obj(200752493:01.@entry,101227190:06)
   {/unl}
[/S]
```

Fig. 1. Tabular and graph representation of the sentence « Je demande pardon aux enfants d'avoir dédié ce livre à une grande personne ».

In the example above, the nine-digit nodes represent UWs. The node "200752493", for instance, replaces the French original "demander", and represents the concept conveyed by the definition "make a request or demand for something to somebody", which corresponds to the synset 752493 of WordNet 3.1, headed by the word "ask". This node is the source node of four different relations: "agt" (= agent), which links "demander" to "je", the speaker, represented by the null UW 00 (modified by the attribute @1, which corresponds to first person of singular); "obj" (= object), which links "demander" to "pardon" (= "forgiveness"), represented by the synset 1227190 ("the act of excusing a mistake or offense" = "forgiveness, pardon"); "gol" (= addressee), which links "demander" to "enfant", represented by the synset 9917593 ("a young person of either sex" = "child, kid, youngster, minor, shaver, nipper, small fry, tiddler, tike, tyke, fry, nestling"), which is modified by two attributes (@def = definite, because the concept was referred to previously, and @pl = plural); and "rsn" (= reason), which links "demander" to the hyper-node ("d'avoir dédié ce livre à un adult" = "for dedicating this book to a grown-up"), represented by the scope 01.

3. Results

The resulting corpus contains the following data:

Entities	Number	
French original	Number of sentences (tokens)	1,684
	Number of sentences (types)	1,601
	Number of words (tokens)	14,304
UNL version	Number of statements	7,008
	Number of UWs (tokens)	13,967
	Number of UWs (types)	1,832
	Number of relations (tokens)	6,971
	Number of relations (types)	36
	Number of attributes (tokens)	10,341
	Number of attributes (types)	1,005

Table 1. Number of entities in Le Petit Prince.

In the table above, the total number of occurrences (tokens) is differentiated from the number of distinct occurrences (types). The French original, for instance, contained 83 repeated sentences, which were UNL-ized only once. As the source document was not lemmatized, we may not provide the number of distinct natural language lemmas, but the whole corpus contained 1,832 different concepts (UWs), and each natural language sentence comprised an average of 4.35 relations. The distribution of the concepts in the four major semantic categories (adjectives, adverbs, noun and verbs) is depicted in the Table 2 below.

Class	Frequency
Adjectives	298
Adverbs	179
Nouns	740
Verbs	615
All	1832

Table 2. Distribution of UWs in different classes.

The most used relations were “obj” (patient), “agt” (agent) and “aoj” (predicative), which were responsible for 52% of the total number of occurrences. 76% of the corpus was described with only 6 relations: “obj”, “agt” and “aoj”, already mentioned; and “mod” (attribute), “man” (manner) and “and” (conjunction). If we add other four relations to this list – i.e., “pos” (possessor), “plc” (place), “tim” (time) and “gol” (addressee) – we would have around 90% of the corpus, as described in Table 3 below.

Relation	Frequency
obj	1498
agt	1088
aoj	1041
mod	819
man	470
and	387
pos	275
plc	241
tim	239
gol	208
qua	145
rsn	101
pur	81
con	70
scn	35
cnt	34
or	30
src	29
bas	28
ben	23
seq	19
ins	18
dur	17
plf	16
equ	15
plt	14
ptn	9
tmf	5
met	3
nam	3
cag	2
coo	2
iof	2
pof	2
per	1
tmt	1

Table 3. Frequency of relations.

This data have been used in a natural language generation project dedicated to the creation of natural language resources (dictionaries and grammars) required for generating the UNL version of Le Petit Prince back onto as many natural languages as possible. This project – LPP – is hosted at the UNLarium and has been addressed by partners and freelancers who have been providing 1) the mappings between UWs and natural language words (NLWs); and 2) the morphological and syntactic information (such as lemma, base form, gender, number, inflectional paradigm, subcategorization frames, etc.) for each NLW. During this process, several issues have been raised, as presented in the next section.

4. Issues

The project Le Petit Prince has been one of the main subjects of the UNL-ization forum hosted at www.unlweb.net/forum, where we have been discussing several problems concerning the interpretation from natural language into UNL. Besides the well-known problems concerning lexical and structural differences between natural languages (the “translation challenges” referred to by Dorr, Jordan & Benoit 1999), there are three main issues that are actually related to the current structure of the UNL:

a) The balance between different levels of representation. In the UNL approach, information conveyed by natural language sentences is represented in one of three possible levels: as nodes in the graph (i.e., as Universal Words); as semantic binary relations linking nodes; or as semantic attributes attached to nodes. This tripartite representation schema has been posing some doubts especially concerning the representation of closed-class categories, such as determiners, prepositions and conjunctions, and some special adverbs. The problems do not affect prototypical words (such as the definite article “le”, to be represented by the attribute .@def; or the conjunction “et”, to be represented by the relation “and”; or the adverb “hier”, to be represented by the UW corresponding to the concept of “yesterday”), but words that express attributes or features that do not fully correspond to existing UNL entities, such as the determiner “certaines” (certain), the preposition “sur” (about), the conjunction “malgré” (despite) or the adverb “bien” (very), which seem not to involve any real independent concept (and should not therefore be represented by UWs), but do not correspond to any existing relation or attribute. In order to avoid this problem, we have been adopting, for the time being, the general procedure that only open lexical classes are represented as UWs, and that all determiners, prepositions and conjunctions that cannot be represented by the current set of relations must be represented as attributes (determiners: @certain, @own, @only, @no, etc.; prepositions: @about, @after, @before, etc.; conjunctions: @although, @unless, etc.). The same happens to the degree adverbs (such as “bien”, “plus”, etc.), which have been represented by degree attributes (@plus, @more, etc.). In this sense, we have been extending considerably the current list of attributes in the UNL Specification, even though we understand that this is a temporary solution for the problem of language-dependent grammar-related words (the updated list of attributes may be found at http://www.unlweb.net/wiki/index.php/Attributes).

b) The interpretation of figures (such as metaphors, hyperboles, ironies, etc.), which are activated by non-explicit background knowledge required for the interpretation of a segment. This happens quite often in the original text, as in “un renard semblable à cent mille autres” (a fox like a hundred thousand others), where the expression “cent mille” (hundred thousand) does not make actual reference to the concept 100,000 but to the idea represented by the attribute @multal. In cases like this, the UNL interpretation of the sentence must not respect the surface structure (i.e., the literal meaning), or we would merely provide a language-dependent representation of the source sentence, which is not the purpose of the UNL approach. Whereas the human interpreter does not have much difficulty in representing the meaning actually conveyed by the original sentence, one of our main commitments is exactly to come up with machine-replicable UNL-ization strategies, i.e., that could be formalized later on in order to replicate the behavior of the human analyzer. For the time being, this is still an open issue, and we have been representing figures of speech by special attributes (such as @metaphor, @hyperbole, and so on), so as to be able to retrieve the most likely candidate forms in natural language generation.

c) The interpretation of anaphora, as in “C’est utile.” (It’s useful.), “C’est étrange.” (It’s strange.) or “Celui-là, se dit le petit prince, tandis qu’il poursuivait plus loin son voyage, celui-là serait méprisé ” (This one, said the little prince, as he continued farther on his journey, he would be scorned). In these cases, the semantic content of the sentence is not saturated yet, and points out to an extrasentential segment, which cannot be referenced by ordinary (intrasentential) nodes. Cross-sentence indexation has been one of the major issues inside the UNL approach – which is sentence-driven – and the need for referring nodes inside external graphs has been leading us to revise the specification, and to overlook natural language sentence boundaries, in order to treat the whole text as a single network (instead of a collection of isolated graphs). This means not only representing relations between sentences and clauses – as in text-driven approaches such as the Rhetorical Structure Theory (Mann & Thompson, 1988) – but reorganizing the whole text structure in a rather concept-driven network, where the idea of sentence plays actually a very small role. The implementation of such strategies has been posing however several issues concerning narrative texts, such as Le Petit Prince, where the order of appearance of the sentences may not be disregarded, under the risk of losing the own narrativity of the text. Accordingly, and for the time being, we have been just creating cross-sentence indexes, which have been represented by the null UW (“00”) along with a specific attribute @ellipsis, so that we can, later on, move back to these cases and propose better indexation mechanisms.

These issues have been leading us to make a comprehensive revision of the current UNL specifications, which has been considered a candidate release for a new standard, the XUNL – the eXtended UNL, which has been the result of a series of discussions about the current limitations and shortcomings of the UNL, especially in comparison to other technologies on knowledge representation and natural language processing. The XUNL is still an ongoing project that has been taking place inside the UNDL Foundation since 2009 as part of the UNL+3 initiative, and whose several innovations have been leading us to claim some independency from the UNL itself, although they share the same basic framework, which is to represent information through semantic networks composed of UWs, relations and attributes.

References

Bonnie J. Dorr , Pamela W. Jordan , John W. Benoit. (1999). A Survey of Current Paradigms in Machine Translation. Advances in Computers, Volume 49, Pages 1–68.

Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

W.C. Mann & S.A Thompson. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3). 243-281.

H. Uchida, M. Zhu, and T. Della Senta. (1999). A gift for a millennium, IAS/UNU, Tokyo, 1999.

H. Uchida, M. Zhu, and T. Della Senta. (1996). UNL: Universal Networking Language - An Electronic Language for Communication, Understanding and Collaboration. UNU/IAS/UNL Center. Tokyo, Japan.