An interlingua aiming at communication on the Web: How language-independent can it be?

Ronaldo Teixeira Martins
ronaldo@nilc.icmsc.sc.usp.br
Lucia Helena Machado Rino
lucia@dc.ufscar.br
Maria das Graças Volpe Nunes
mdgvnune@icmc.sc.usp.br
Gisele Montilha
gisele@nilc.icmsc.sc.usp.br
Osvaldo Novais de Oliveira Jr.
chu@if.sc.usp.br

Núcleo Interinstitucional de Lingüística Computacional (NILC/São Carlos)
http://nilc.icmsc.sc.usp.br
CP 668 – ICMC-USP, 13560-970 São Carlos, SP, Brazil

Abstract

In this paper, we describe the *Universal Networking Language*, an interlingua to be plugged in a Web environment aiming at allowing for many-to-many information exchange, 'many' here referring to many natural languages. The interlingua is embedded in a Knowledge-Base MT system whose language-dependent modules comprise an encoder, a decoder, and linguistic resources that have been developed by native speakers of each language involved in the project. Issues concerning both the interlingua formalism and its foundational issues are discussed.

1. Introduction

The widespread use of the Web and the growing Internet facilities have sparked enormous interest in improving the ways people use to communicate. In this context multilingual Machine Translation systems become prominent, for they allow for a huge information flow. To date, MT systems have been built under limited conditions, of which we highlight two: i) in general, they mirror one-to-many(languages) or many(languages)-to-one approaches, often involving English at the "one" end; ii) communication is reduced to basic information exchange, ignoring richness and flexibility implied by human mind. The first limitation has been seldom overcome, since it requires a robust environment and research teams that can cope with knowledge of several languages[1], to derive precise automatic language analyzers and synthesizers. The second limitation follows up the first: adding up communicative issues to linguistic processing/modeling makes still harder to overcome MT limitations.

In this article, we elaborate on work using an interlingua conceived to overcome the first limitation, i.e., to allow for a many-to-many information exchange environment, which shall be plugged in a nontraditional Internet platform. The goal is to allow interlocutors to entangle communication even if they do not share the same mother tongue or the English

[1] Standing, most often, for natural language, or NL.

language, unlike MT systems that have just one language at one of their edges. As the main component of a Knowledge-Base MT system (hereafter, KBMT), the interlingua approach has been developed under the Universal Networking Language Project, or simply UNL Project. What makes the interlingua UNL special is its intended use: as an electronic language for networks, it has to allow for high quality[2] conversation systems involving many languages. As the main component of a KBMT system, it has to be sufficiently robust to ground research and development (R&D) of the language-specific modules to be attached to the system. It is this latter perspective that is undertaken here: from the viewpoint of R&D, we discuss how broad, or language-independent, the interlingua UNL is, especially focusing on its syntax and coverage. In addition to being consistent and complete to represent meaning, we also consider its sharing by researchers all around the world, which is an important bottleneck of the UNL Project, since information exchange by researchers during R&D brings about the problems introduced by the interlingua UNL itself, concerning both its formalism and foundational issues. Before discussing this topic in Section 5, we present an overview of the UNL Project (Section 2) and describe the main features of the interlingua UNL (Section 3). In Section 4, we describe the UNL system architecture. Hereafter, 'interlingua UNL' will be simply referred to as UNL, the acronym for *Universal Networking Language*. Also, the viewpoint presented here is that of interlingua users who experience R&D for a given NL, and not of its authors.

2. The UNL Project

The UNL Project[3] has been launched by the United Nations University to foster and ease international web communication by means of NLP systems. Its main strength lies on the development of the UNL, as a unique semantic (or meaning) representation that can be interchanged with the various languages to be integrated in the KBMT system. In the UNL Project, plug-in software to encode NL texts onto UNL ones (NL-UNL encoders) and to decode UNL into NL texts (UNL-NL decoders) have been developed by R&D groups in their own native languages. The modules to process Brazilian Portuguese[4], for example, have been developed by a team of Portuguese native speakers that comprises linguists, computational linguists, and computer experts. Such packages will be made available in WWW servers and will be accessible by browsing through Internet, thus overcoming the need for people all around the world to learn the language of their interlocutors. Several linguistic groups have signed to the Project, namely: the Indo-European (Portuguese, Spanish, French, Italian, English, German, Russian, Latvian and Hindi), the Semitic (Arabic), the Sino-Tibetan (Chinese), the Ural-Altaic (Mongolian), the Malayan-Polynesian (Indonesian), and the Japanese.

On the one hand, the main strength of the Project is that knowledgeable specialists address language-dependent issues of their mother tongue, most of which are related to R&D of the encoding and decoding modules and to the specification of the NL-UNL lexicon. On the other hand, this also represents a crucial problem faced by the project participants, for distinct groups may interpret the interlingua specification differently. There is thus the need for a consensus about the UNL formalism,

[2] By 'high quality' we mean 'at least allowing for readability and understandability by any user'.

[3] A description of both, the Project and the UNL itself, can be found in http://www.unl.ias.unu.edu/.

[4] Hereafter referred to as Portuguese or by its acronym, BP.

bringing about an assessment of its coverage, completeness, and consistency, all features that will be discussed shortly.

3. The Universal Networking Language

The UNL is a formal language designed for rendering automatic multilingual information exchange. It is intended to be a cross-linguistic semantic representation of NL sentence meaning, being the core of the UNL System, the KBMT system developed by H. Uchida (1996) at the Institute of Advanced Studies, United Nations University, Tokyo, Japan.

UNL subsumes a tridimensional theory of (sentence) meaning, whose components are defined according to one of the following sets (Martins et al., 1998a): concepts (e.g., "cat", "sit", "on", or "mat"), concept relations (e.g., "agent", "place", or "object"), and concept predicates (e.g., "past" or "definite"). Such components are formally and correspondingly represented by three different kinds of entities, namely: Universal Words (UWs), Relation Labels (RLs), and Attribute Labels (ALs). According to the UNL syntax, information conveyed by each sentence can be represented by a hypergraph whose nodes represent UWs and whose arcs represent RLs. To make symbol processing simpler, hypergraphs are often reduced to lists of ordered binary relations between concepts, as it is shown in Figure 1 for the sentence (1) *The cat sat on the mat.*[5]

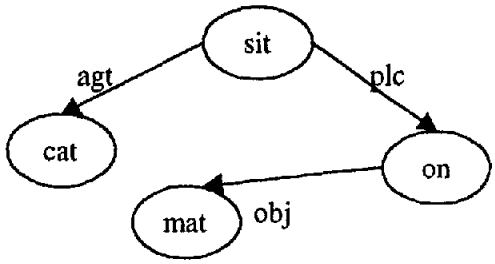

Figure 1a: UNL hypergraph representation of the English sentence "The cat sat on the mat"

```
agt(sit.@entry.@past,cat.@def)
plc(sit.@entry.@past,on)
obj(on,mat.@def)
```

Figure 1b: UNL linear representation of the English sentence "The cat sat on the mat."

UWs are labels for concept-like information, roughly corresponding to the lexical level in the sentence structure. They comprise an open large inventory, virtually capable of denoting every non-compositional meaning to be conveyed by any speaker of any language. For the sake of representation, these atomic semantic contents are associated to English words and expressions, which play the role of semantic labels. However, there is no one-to-one mapping between the English vocabulary and the UNL lexicon, for UNL, as a multilingual representation code, is larger than the English vocabulary. To avoid unnecessary proliferation of the UNL vocabulary and to certify that standards be observed by UNL teams, control over the specification of the UW set is centered at the UNL Center, in Japan.

Several semantic relationships hold between UWs, namely synonymy, antonymy, hyponymy, hypernymy and meronymy, which compose the UNL Ontology. Steady semantic valencies (such as agent and object features) can also be represented, forming the UNL Knowledge-Base. Both Ontology and Knowledge-Base aim at constraining the scope of UW labels, whenever ambiguity is to be avoided. The UNL representation of sentence (1), for example, can be ambiguous

[5] 'sit', 'cat', 'on' and 'mat' are UWs; 'agt' (agent), 'plc' (place) and 'obj' (object) are RLs; '@def', '@entry' and '@past' are ALs.

in Romance languages, for the translation of 'cat' should make explicit the animal sex: if male, it would be "gato" (Portuguese and Spanish), "gatto" (Italian), "chat" (French), whereas different names would have to be used for the female cat. Instead of having a unique UW 'cat', it is thus quite feasible to have a whole structure in which 'cat' is only the hyper-ordinate option.

For the English-UNL association not to undermine the intended universality of the UW inventory, its semantic-orthograpical correspondence has to be considered rather incidental, or even approximated. It is not always the case that extensions[6] of a UW label and of its corresponding English word coincide. The extension of the English word "mat", for example, does not exactly coincide with the extension of any Portuguese word, although we can find many overlaps between "mat" and, e.g., "capacho" (Portuguese). Portuguese speakers, however, would not say "capacho" for the ornamental dishmat, as would not English speakers use the word "mat" for a fawner (still "capacho" in Portuguese). Since each language categorizes the world in a very idiosyncratic way, it would be misleading to impose a straightforward correspondence between lexical items of two different languages. In UNL, this problem has been overcome by proposing a rather analogic lexicon, instead of a digital one. Although discrete, UWs convey continuous entities, in the sense that semantic gaps between concepts are fulfilled by the UNL Knowledge-Base, as it is shown for the UW 'mat' in Figure 2. Granularity thus plays an important role in UNL lexical organization and brings flexibility into cross-linguistic lexical matching.

[6] Cf. (Frege, 1892), extension here is used to establish the relationship between a word and the world, opposed to intension, referring to the relationship between a word and its meaning.

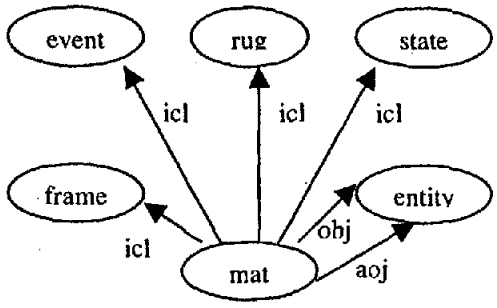

Figure 2a: UNL hypergraph partial representation for the meaning denoted by the English word "mat"

"mat"
"mat(aoj>entity)"
"mat(icl>event)"
"mat(icl>frame)"
"mat(icl>rug)"
"mat(icl>state)"
"mat(obj>entity)"

Figure 2b: UNL partial linear representation for the meaning denoted by the English word "mat"

While lexical representation in UNL comprises a set of universal concepts signaled by UWs, the cross-lexical level involves a set of ordered binary relations between UWs, which are the Relation Labels (RLs). RLs specification are similar to Fillmore's semantic cases (1968), with RLs corresponding to semantic-value relations linking concept-like information. There are currently 44 RLs, but this set has been continuously modified by empirical evidence of lack, or redundancy, of relations. The inventory of RLs can be divided into three parts, according to the functional aspects of the related concepts: ontological, event-like and logical relations. Ontological relations are used as UW constraints in reducing lexical granularity or avoiding ambiguity as shown above, and they help positioning UWs in a UNL lexical structure. Five different labels are used to convey ontological relations: icl (hyponymy), equ (synonymy), ant (antonymy), pof (meronymy), and fld (semantic field).

UNL depicts sentence meaning as a fact composed by either a simple or a complex event, which is considered here the starting point of a UNL representation, i.e., its minimal complete semantic unit. Event-like relations are assigned by an event external or internal structure, or by both. An event external structure has to do nearly always with time and space boundaries. It can be referred to by a set of RLs signaling the event co-occurrent meanings, such as[7] its environment (scn); starting place (plf), finishing place (plt), or, simply, place (plc); range (fmt); starting time (tmf), finishing time (tmt), or, simply, time (tim); and duration (dur). Action modifiers, such as manner (man) and method (met) can also qualify this structure. An event internal structure is associated to one of the following simple frames: action, activity, movement, state, and process, each expressing different RLs in the event itself, including its actors and circumstances.

Event actors are any animate or inanimate character playing any role in events, which can be the main or the coadjutant actors. There can be up to eight actors, signaled by the following RLs: agent (agt), co-agent (cag), object (obj), co-object (cob), object place (opl), beneficiary (ben), partner (ptn) and instrument (ins). They can also be coordinated through the RLs conjunction (and) and disjunction (or), or subordinated to each other by possession (pos), content (cnt), naming (nam), comparison (bas), proportion (per), and modification (mod). They can still be quantified (qua) or qualified by the RLs "property attribution" (aoj) and co-attribution (cao). It is possible to refer to an "initial actor" (src), a "final actor" (gol), or an "intermediary actor" (via). Finally, spatial relationships can also hold between actors: current place (plc), origin (frm), destination (to), and path (via). Besides single events, there can still be complex cross-event relationships which express either paralleled events – co-occurrence (coo), conjunction (and), and disjunction (or) – or hierarchically posed events – purpose (pur), reason (rsn), condition (con), and sequence (seq). They can all be referred to as logical relations, since they are often isomorphic to first-order logic predicates.

According to the UNL authors, it is possible to codify any sentence written in any NL into a corresponding UNL text expressing the sentence meaning through the use of the above RLs. This is still a claim to be verified, since cases of superposition and competition between different RLs have been observed, as it is discussed in Section 5.

In addition to UWs and RLs, UNL makes use of predicate-like information, or Attribute Labels (ALs), which are names for event and concept "transformations", in a sense very close to that intended by Chomsky (1957, 1965). They are not explicitly represented in a UNL hypergraph, although they are used to modify its nodes. ALs can convey information about concept intensions and extensions. In the former case, ALs name information about utterers' intensions over either specific parts of a sentence (focus, topic, emphasis, theme) or the whole structure (exclamation, interrogation, invitation, recommendation, obligation, etc.). In the latter case, ALs refer to spatial (definite, indefinite, generic, plural) or temporal (past, present, future) information, or still, temporal external (begin-soon, begin-just, end-soon, end-just) or internal (perfective, progressive, imperfective, iterative) structures. To differentiate ALs from UWs, ALs are attached to UWs by the symbol ".@". The concept expressed by the UW 'sit' in "sit.@entry.@past", for example, is taken as the starting point (.@entry) of the corresponding hypergraph and it is to be modified by temporal information (.@past).

[7] RLs names are bracketed.

4. The UNL System

The UNL system architecture consists of two main processes, the encoder and decoder, and several linguistic resources, each group of these corresponding to a NL embedded in the system, as depicted in Figure 3.

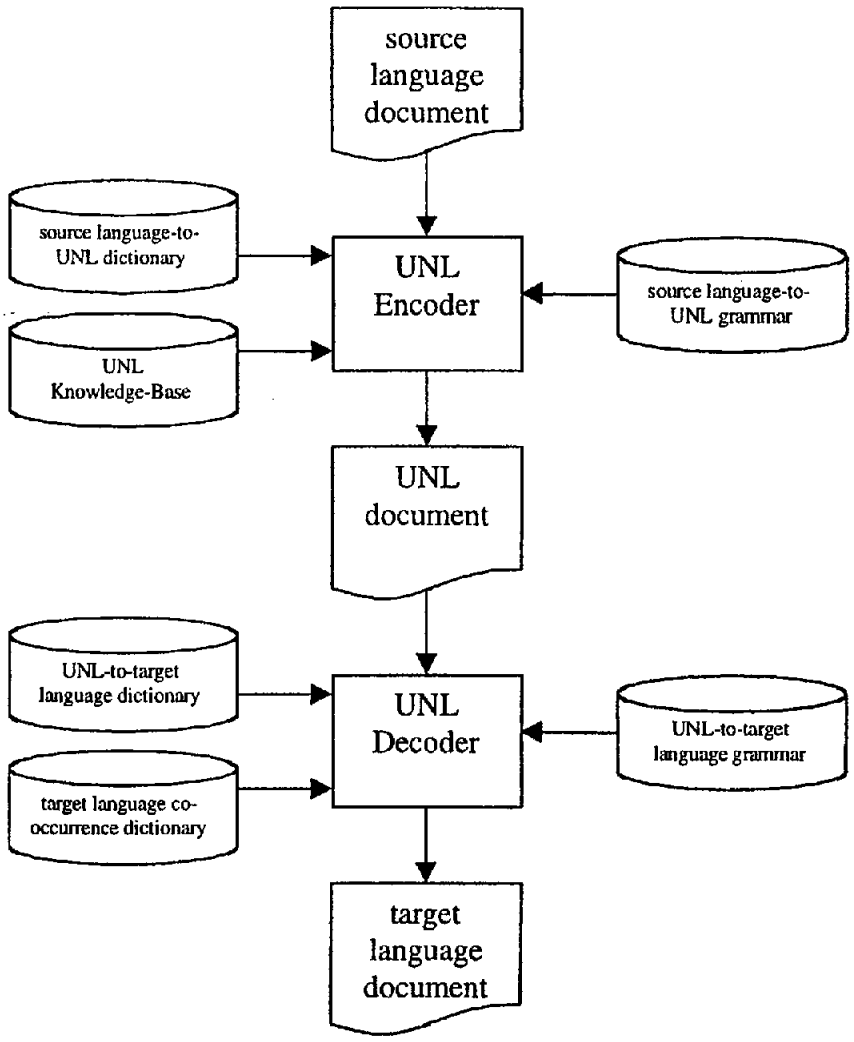

Figure 3: The UNL System Architecture

A source document (SLD) conveys written text on any subject, in any of the NLs considered. There is no constraint in the domain or structure of the SLD, but there is necessarily a loss of semantic expressiveness during NL-UNL encoding. The goal of the UNL is not, in principle, to fully preserve text meaning, but only its main components, i.e., those considered to be essential. However, there is no measurable account as to what is essential in the UNL Project. By convention, this is linked to what has been called the literal meaning, which is directly derived from interpreting the sentence surface structure. Therefore, there is no room to represent content that is not directly mapped onto the NL syntactic-semantic licensed structures.

The NL-UNL encoding tool, or UNL Encoder, is generic enough to handle all the

languages included in the Project. Apart from the (supposedly) universal knowledge-base, used to fill-in possible interlexical gaps when mapping is not precise, all other linguistic resources are language-dependent. The source grammar essentially guides the elicitation of the sentence semantic structure into its corresponding UNL structure, by determining RLs and ALs, always giving priority to information content.

The UNL-NL decoding tool, or UNL Decoder, works in the opposite way to the Encoder. Besides the lexicon and the grammar, a cooccurrence dictionary is also used at this stage, to disentangle lexical choice. The target grammar is responsible for the semantic-syntactic mapping, now resolving semantic organization by making syntactic and dependence choices between UWs, taking RLs and ALs into account.

5. Remarks on language-independence

The main strength of the UNL Project rests on human expertise: language-specific aspects to be included in the multilingual KBMT system are handled by native speakers of that language, in an attempt to overcome the need of representing knowledge across several languages or cultures. It has been successful in developing NL-driven resources and processes by researchers all around the world. For example, the BP UNL lexicon has over 65,000 entries that are categorized according to grammatical and some semantic features, and this will be extended considerably in the future to cover the Portuguese vocabulary to a greater extent. Up to the present time, only decoding systems customized to each NL have been plugged into a general decoder skeleton (provided by the UNL Center) and have already been assessed, producing promising results. The BP decoder, for example, is able to produce outputs whose literal meaning is preserved in most cases (Martins et al., 1998b), using handcoded UNL expressions. Actually, to decode any UNL text, NL-UNL encoding has to be handmade, since customization of the UNL Encoder to each NL has not yet been undertaken in the project. In spite of the promising decoding results, a) output quality varies enormously with UNL sentences encoding, which can be different across distinct research groups; b) communicative aspects of information exchange on the web are not explored in depth, as it can be seen through the list of RLs or ALs. UNL is not knowledge intensive and there are no guidelines as to consistently recognize or extract such kind of information from the surface of the source texts.

There are several reasons why interpretation and use of the UNL among the various teams are not uniform, including cultural aspects and syntax differences of the languages involved. Using English as the *lingua franca* for communication and cooperation among the research groups and as the knowledge representation language has also brought limitations into the Project, since it implies a non-desirable level of language-dependence. This is inevitable, however, for limitations definitely come along with the choice made. For example, attaching a NL word to a UW may be difficult, owing to the cross-references introduced by using English to convey UNL symbols. Resuming the example shown in Figure 1, this is the case of the UW "on" in (1b): the preposition 'on' fills in the position feature of the verb 'sit' and, thus, is represented in UNL correspondingly as the second term of the binary relation 'plc' and the first term of 'obj'. This, undoubtedly, is critical, for 'sit' can be juxtaposed to other prepositions leading to different meanings, which, in turn, may introduce different sets of binary relations, implying a high-level complexity in the UNL representation. As a result, languages whose syntactic structures deeply differ from the English ones may

present an additional level of complexity that makes mapping to/from UNL impossible or unrealistic. In this respect, we have not been facing many problems in fitting Portuguese structures with UNL ones, since Portuguese, like English, is an inflectional language that also employs prepositional constructions. However, prepositions in Portuguese may play considerably different roles compared to English. Various extensions of the English spatial prepositions "on", "over" and "above", for example, are subsumed in Portuguese by a single form "sobre" (which may also mean *about*). Therefore, in Portuguese, cats could be, at the same time, not only "on" but also "over" and "above" mats. Only world knowledge, associated to contextual indexes, both absent in the referred UNL hypergraph, could avoid the unsuited encodings *The cat sat over the mat.* or *The cat sat above the mat.* from the Portuguese sentence "O gato sentou *sobre* o tapete".

Another problem related to the sentence *The cat sat on the mat.* refers to the existence of competing analyses: it is quite plausible that a UNL representation suggesting a noun phrase instead of a full sentence holds for this sentence. It so happens when the arc between 'sitting' and 'cat' concepts are labeled by the RL 'obj', instead of the RL 'agt' in (1), as it is shown in Figure 1a', yielding the UNL text shown in Figure 1b'.

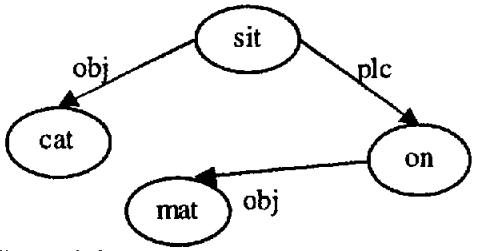

Figure 1a': UNL hypergraph representation of the English sentence "The cat sat on the mat."

```
obj(sit.@entry.@past,cat.@def)
plc(sit.@entry.@past,on)
obj(on,mat.@def)
```

Figure 1b': UNL linear representation of the English sentence "The cat sat on the mat."

Both analyses are equally accurate and can lead to good NL surface expressions, although they refer to different semantic facts. Indeed, to define an object relationship between "sitting" and "cat" is to say that the cat was already sat before the beginning of the event (e.g., *The cat sat on the mat ate the fish.*). In this case, the animal does not actually perform the action, but is conditioned to it, the main performer position being empty, thus yielding the referred noun phrase. In Figure 1, instead, the cat on its own has taken the sitting position, therefore introducing an agent relationship. These two different semantic facts may correspond, in English, to a single surface structure. Indeed, (1) is orthographically identical to (1'). However, other languages (e.g., Portuguese) do behave differently.

Although it is also possible to have, in Portuguese, the same surface structure corresponding to both UNL representations ("sentado no tapete"), it is more feasible to have, for each case, completely different constructions. In the case depicted by Figure 1, the UW "sit" would be associated to the verb "sentar" (corresponding to "to sit"). Thus, the generation result should be something like "O gato sentou no tapete" or "O gato sentado no tapete". On the other hand, for Figure 1', the same UW 'sit' would be generated in a completely different way, corresponding to the passive form of the Portuguese expression "colocar sentado" (*to be put in a sitted position*), for which there is no adequate English surface expression.

Distinguishing such situations to cope with syntactic-semantic troublesome mappings, though interesting, is a highly

context-sensitive task, often surpassing sentence boundaries. UNL descriptions do not address such fine-grained level of meaning representation, being limited to meanings derived from context-free source sentences, even when context-freeness implies insufficient information. When this is not possible, UNL offers a default analysis for semantically ambiguous sentences, in which case we can say that the UNL representation is probabilistic, rather than deterministic.

The way we believe some of UNL limitations can be overcome and/or minimized is by designing a fully-fledged testing procedure to assess outputs of both decoder and encoder for the various languages. Since the same encoding and decoding procedures have been delivered to the UNL teams, it is possible that part of the set of rules or translation strategies of a given team may be interchangeable with another one from a different language. In this way, sharing procedures may become a warranty for common ground assessment of the varied models, in which case it may be possible to make eligible concurrent strategies equally available for the languages involved.

Concerning the UNL means to disambiguate or proceed to reference resolution or other discourse figures, most of the troublesome occurrences are enclosed in the treatment issued by specialists and, thus, they are constrained to, and handled by, at the level of native speakers use. This measure can be somewhat fruitful, provided that each signatory of the Project finds a way to trace a UNL text back onto its own NL text or vice-versa, making a proper use of the UNL syntax or symbols. This, in fact, can be a good method to evaluate (de)coding: once a UNL code has been produced from any NL text, this code can be the source to decoding into the same NL, in order to compare the original NL text with the automatically generated one. Evaluation, in this case, can be carried out by the same research group responsible for both processes.

Compared to other interlingua approaches (e.g., Mikrokosmos, Gazelle, or Kant), the UNL Project is in a much earlier stage – most of those are over 10 years old, while the UNL one is about 3 years old – but it is much more ambitious than most of the current systems under construction. For UNL is actually a front-end to a many-to-many communication system, with no constraints that are normally inherent in MT systems. Since knowledge is specified by native speakers for each NL module, grammar, semantics and world knowledge can be well founded. Its limitations, from a conceptual viewpoint, are shared by most of its counterparts, as in treating text at the sentence level only. In addition, by no means is the UNL system committed to event replication as it is the case of human translation. Automatic strategies have no psychological motivation whatsoever and are solely based upon computer efficiency principles, namely time and space.

Acknowledgments

The development of resources for Brazilian Portuguese in the UNL Project has been sponsored by the Institute of Advanced Studies of the United Nations University. The authors are also grateful to CNPq and Finep (Brazil) for the financial support and to Mr. Tadao Takahashi, the coordinator of the Brazilian branch in the UNL Project.

References

Chomsky, N. (1957). *Syntactic Structures.* The Hague, Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT Press, Cambridge, MA.

Fillmore, C. (1968). The case for case. In Bach, E. and Harms, R.T. (orgs.), *Universals in linguistic theory*, pp. 1-88. Rinehard and Winston, New York.

Frege, G. (1892). On Sinn and Bedeutung. In Beaney, M. (ed.), *The Frege Reader*. Blackwell Publishers, Malden, MA, 1997.

Martins, R.T., Rino, L.H.M., Nunes, M.G.V. (1998a). *As Regras Gramaticais para a Decodificação UNL-Português no Projeto UNL*. Relatório Técnico 67. Instituto de Ciências Matemáticas e da Computação. Universidade de São Paulo, São Carlos.

Martins, R.T.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N. (1998b). *Can the syntactic realization be detached from the syntactic analysis during generation of natural language sentences?* III Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'98). Porto Alegre - RS. Novembro.

Uchida, H. (1996). *UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration*. UNU/IAS/UNL Center. Tokyo, Japan.