The Workshop Programme

9:00 - 9:45	Hiroshi Uchida. The UNL: A language for computers - How to develop a language for computers -. *Invited talk*
9:45 – 10:00	**BREAK**
10:00 – 10:30	Issues in Generating from Interlingua Representations. Stephan Busemann
10:30 – 11:00	The UNL distinctive features: evidences through a NL-UNL encoding task. Ronaldo Teixeira Martins, Lúcia Helena Machado Rino
11:00 – 11-15	**COFFEE BREAK**
11:15 – 11:45	Structural and Lexical Transfer: From an UNL graph to an Equivalent NL Dependency Tree. Etienne Blanc, Gilles Sérasset, WangJu Tsai
11:45 – 12:15	Some Lexical Issues of UNL. Igor Boguslavsky.
12:15 – 12:45	A rationale for using UNL as an Interlingua and more in various domains. Christian Boitet.
12:45 – 13:15	UNL, Challenges and misunderstanding. Some answers. Jesús Cardeñosa, Edmundo Tovar
13:15 – 13:30	**DEBATE**

Workshop Organisers

Edmundo Tovar. UNL-Spanish Language Centre; Validation and Business Applications Group, Universidad Politécnica de Madrid, Spain. E-mail: etovar@fi.upm.es

Carolina Gallardo. UNL-Spanish Language Centre; Validation and Business Applications Group, Universidad Politécnica de Madrid, Spain. E-mail: carolina@opera.dia.fi.upm.es

Workshop Programme Committee

Jesús Cardeñosa. UNL-Spanish Language Centre; Validation and Business Applications Group Universidad Politécnica de Madrid, Spain.
E-mail: carde@fi.upm.es

Igor Boguslavsky. Institute for Information Transmission Problems, Russian Academy of Sciences, Russia. E-mail: bogus@iitp.ru

Christian Boitet.. Universite' Joseph Fourier, GETA- CLIPS, Grenoble, France.
Email: Christian.Boitet@imag.fr

Irina Prodanof. Institute of Computational Linguistics Consorzio Pisa Riserche-Settore Linguistica, Italy. E-mail: irina@ilc.pi.cnr.it

Table of Contents

Author Index

Issues in Generating Text from Interlingua Representations

Stephan Busemann

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
busemann@dfki.de

Abstract

Multi-lingual generation starts from non-linguistic content representations for generating texts in different languages that are equivalent in meaning. In contrast, cross-lingual generation is based on a language-neutral content representation which is the result of a linguistic analysis process. Non-linguistic representations do not reflect the structure of the text. Quite differently, language-neutral representations express functor-argument relationships and other semantic properties found by the underlying analysis process. These differences imply diverse generation tasks. In this contribution, we relate multi-lingual to cross-lingual generation and discuss emergent problems for the definition of an interlingua.

1. Introduction

In this contribution, we relate multi-lingual to cross-lingual generation and discuss emerging problems for the definition of an interlingua. Multi-lingual generation starts from non-linguistic content representations for generating texts in different languages that are equivalent in meaning. The generation of weather forecasts or environmental reports are typical examples. In contrast, cross-lingual generation is based on a language-neutral content representation which is the result of a linguistic analysis process. Generation for machine translation is a most prominent example.

Non-linguistic representations do not specify linguistic semantics nor do they reflect the structure of the text to be generated. In contrast, language-neutral representations express functor-argument relationships and other semantic properties found by the underlying analysis process. These differences imply diverse generation tasks.

However, there are also commonalities. In both cases, generation is the mapping of some semantic representation onto linguistic strings. We may assume a single generation process that uses different separately defined language-specific knowledge sources. In both cases, we may view the underlying representation as an interlingua, since it attempts to cross the language barrier by providing content descriptions independently of the target language.

An instance of each type of tasks has been implemented using the generation system TG/2 (Busemann, 1996), quickly overviewed in Section 2.. The usage of the same framework allows us to relate the tasks to each other (Section 3.) and to gain insights relevant to a coherent definition of interlinguas, generation tasks, and generation knowledge (Section 4.).

2. TG/2 in a Nutshell

TG/2 is a flexible production system that provides a generic interpreter to a set of user-defined condition-action rules representing the generation grammar. The generic task is to map an input structure onto a chain of terminal elements as prescribed by the rule set. The rules have a context-free categorial backbone used for standard top-down derivation, which is guided by the input representation. The rules specify conditions on input (“tests”) determining their applicability and allow navigation within the input structure (“access functions”).

The right-hand side of a rule can consist of any mixture of terminal elements (canned text) or other categories associated with an access function. The presence of canned text is useful if the input does not express explicitly everything that should be generated. With very detailed input, the terminal elements of the grammar will usually be words.

Given a category C and some (piece of) input structure I, production rules are applied through the standard three-step processing cycle:

1. Identify the applicable rules;
2. Select a rule on the basis of some (freely programmable) conflict resolution mechanism; and
3. Apply that rule.

A rule is applicable if its left-hand side category is C and its tests hold on I. A rule is applied by processing its right-hand side elements from left to right. Canned text is output right away, and non-terminal elements induce a new cycle with the new category and the return value of the access function. Processing terminates when all right-hand side elements have been realized successfully. In the case of a failure, processing backtracks to step 2. If no more rules are applicable, a global failure occurs. For details see (Busemann, 1996).

3. Relating Two Distinct Generation Tasks

TG/2 has been used in a variety of NLG tasks. We look at multi-lingual report generation and cross-lingual summarization. We then locate the tasks on a scale ranging from shallow to in-depth generation, and discuss advantages and drawbacks of these locations.

3.1. Task 1: Generating air quality reports from measurement data

Reports about air quality in a German-French border region (Busemann and Horacek, 1998) are currently

```
[(COOP THRESHOLD-PASSING)
 (LANGUAGE ENGLISH)
 (TIME [(PRED SEASON)
        (NAME [(SEASON WINTER)
               (YEAR 2001)])])
 (POLLUTANT SO2)
 (SITE "Saarbruecken-City")
 (SOURCE [(THRESHOLD-TYPE MIK-WERT)])
 (EXCEEDS [(STATUS YES) (TIMES 1)])]
```

Figure 1: A Non-Linguistic Input Expression for Report Generation: "In Winter 2001 at the measuring station at Saarbrücken-City, the MIK value for sulfur dioxide was exceeded once."

produced in six languages (a web demo is available at `http://www.dfki.de/service/nlg-demo`). The reports are based on real measurement data taken from a database and on the user's parameters determining the type of the report (time series, average or maximum value description, threshold passing description). A report consists of up to six statements most of which are verbalized by TG/2. The initial text organization stage retrieves the relevant data, decides about the content of the statements and defines their order. For each statement to be verbalized by TG/2 it produces a domain-oriented non-linguistic intermediate feature structure serving as input to TG/2 (cf. Figure 1 for an example). Input expressions for TG/2 may specify e.g. the pollutant, the actual measurements, and their date and location. Moreover, further information is specified according to the user's choice of parameters. It should be noted that some input is just carried forward from the original system input (in Figure 1, this is `LANGUAGE, TIME, POLLUTANT, SITE, THRESHOLD-TYPE`), whereas other information originates from the DB query and text organization stage (`COOP` and `EXCEEDS` in Figure 1).

The text organization stage is entirely content-oriented, and the intermediate feature structures do not exhibit linguistic properties. The 'language' feature causes the selection of the rule set for the language requested. The determination of linguistic structure for each input expression is achieved by the TG/2 grammar rules. Since implicit information is associated with some parts of input expressions, canned text is used to make it explicit at the surface. An example in Figure 1 is the added notion of "at the measuring station at" in the case of `(SITE "Saarbrücken-City")`, which is verbalized through the rule in Figure 2.

The grammars comprise about 100-120 rules for each language and are specifically designed for this application. The development of a grammar for another language takes between one and three weeks depending on skills.

3.2. Task 2: Generating medical scientific text for summaries

This generation task occurred in the context of the cross-lingual text summarization system MUSI (Lenci et al., 2002). MUSI involves a combination of analysis and generation similar to machine translation. An interlingua approach was chosen to represent selected English and Italian medical scientific sentences in a language-neutral way. The sentences can be complex and quite long (50 words are no exception). Interlingua expressions were fed to sentence generation components producing the elements of a French or German summary.

```
(defproduction site "S01"
  (:PRECOND
     (:CAT SITE-E
      :TEST ((always-true)))
   :ACTIONS
     (:TEMPLATE
        "at the measuring station at "
        (:RULE SITE-NAME-E (self)))))
```

Figure 2: Making Implicit Meaning Explicit: A TG/2 grammar rule. The rule is "unconditioned" and uses the current piece of input structure to access the site name.

The generation of German sentences (Busemann, 2002) starts from so-called IRep4 interlingua expressions. A sample IRep4 expression is shown in Figure 3. IRep4 expressions are hierarchical predicate-argument structures complemented by a rich variety of features and modifiers. The basic elements are atomic and predicative concepts, forming an ontology shared across the MUSI system. In particular, predicative frames are based on the SIMPLE formal specifications (Lenci et al., 2000). IRep4 expressions are composed of `PROP` and `ITEM` elements used to represent propositions and terms, respectively. Although IRep4 is in principle a semantic representation language, its expressions also keep track of some syntactic properties of the source language elements. For instance, number and determiner information is specified for NPs as well as categorial information for propositions (`CAT`). This information can be very useful in guiding text generators.

IRep4 is suitable for representing the semantics of very complex sentences, but at the same time, it leaves room for various degrees of specification. In fact, co-reference resolution, attachment ambiguities and the incorrect identification of arguments and modifiers are common sentence analysis problems that may lead to incomplete output. To cope with these problems, IRep4 has been designed to integrate possibly underspecified or fragmentary representations. This feature greatly enhances the robustness of the system and can guarantee a better interface with the text analysis component.

A direct interpretation of IRep4 by TG/2 would require choosing the lexemes and the syntactic realizations. This could have been achieved within the TG/2 grammar through complicated tests. These choices partly depend on each other, which would have caused massive backtracking. Moreover, testing the presence of a concept in IRep4 would have been triggered by rules expanding the syntactic category of the lexemes (part of speech), e.g. the rule `Noun` → `"acetylcholin"` would have been associated with a test whether the current concept was `C_acetylcholine`. As there would have been hundreds of these, concerns of processing efficiency were in order. Finally, a pre-existing grammar should be reused that was not previously adapted

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                        Value = C_level;
                                        Det = DEF;
                                        Mod1 = [RESTR, ITEM{
                                                Value = C_sight;
                                                Number = PLUR; Det = DEF;
                                                Mod1 = [RESTR, C_muscarinic];
                                                Mod2 = [RESTR, ITEM{
                                                        Value = C_substance;
                                                        Number = PLUR;
                                                        Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

Figure 3: IRep4 Expression for “Die Wirkungen werden durch einen kompetitiven Antagonismus zu Acetylcholin auf dem Niveau der muskarinischen Bindungsstellen dieser Substanzen verursacht.” [The effects are caused by a competitive antagonism with acetylcholine on the level of the muscarinic sights of these substances.].

to IRep4.

For these reasons it appeared more convenient to introduce an initial sentence planning stage. The resulting representation – see Figure 4 for an example corresponding to Figure 3 – forms the input to TG/2. It can be viewed as a syntactically enriched, language-specific paraphrase of the underlying IRep4 expression. It represents explicitly the linguistic structure of the sentence. The TG/2 grammar is responsible for word order and inflection. Very much like in a classical sentence realization system, no canned text parts are used. If a phrase like “at the measuring station at” had to be generated here, an underlying interlingual semantic expression would be mandatory.

A pre-existing TG/2 grammar for German syntax was reused and adapted to the needs of MUSI (Busemann, 2002; Lenci et al., 2002). Its final version comprises over 950 rules.

3.3. Shallow and in-depth generation

The notion of shallow generation, as opposed to in-depth generation, has been coined by (Busemann and Horacek, 1998) to describe a distinction corresponding to that of shallow and deep analysis. In language understanding deep analysis attempts to “understand” every part of the input, while shallow analysis tries to identify only parts of interest for a particular application, omitting others. In-depth generation is inherently knowledge-based and theoretically motivated, whereas shallow generation quite opportunistically models only the parts of interest for the application in hand. Often such models will turn out to be extremely shallow and simple, but in other cases much more detail is required. Thus, techniques such as those developed within TG/2 for varying modeling granularity according to the requirements posed by the application are a prerequisite for reusing NLG systems.

Obviously a shallow NLG system is, in general, based on representations that carry implicit meaning. We call this shallow input. Additional text has to be “invented” by the generator (in TG/2, this is usually achieved using canned text in the grammar).[1] This leads to domain-dependent, shallow grammars that cannot be reused easily for another task. The in-depth models assume a very fine-grained grammar describing all the linguistic distinctions covered by the interlingua. Such a grammar corresponds closely to familiar generic linguistic resources.

The report generation task described was solved by a typical shallow approach, whereas the MUSI generation task required an in-depth model.

The tension between shallow and in-depth generation has been discussed further in the literature. According to Reiter and Mellish, shallow techniques (which they call “intermediate”) are appropriate as long as corresponding in-depth approaches are poorly understood, less efficient, or more costly to develop (Reiter and Mellish, 1993). Bateman and Henschel describe ways of compiling specialized grammars out of general resources (Bateman and Henschel, 1999). A platform for generating, storing and reusing representations is described in (Calder et al., 1999), showing that such reuse can be seen as a shallow methodology to text generation. A major conclusion seems that there is no dichotomy between both approaches, but that shallow systems can indeed be based on theoretically sound in-depth models.

In practice though, NLG tasks turn out to be highly diverse, and no NLG system could be reused for a new application off the shelf. The necessary effort for adaptation and extension of large existing in-depth resources such as KPML (Bateman, 1997) or FUF/Surge (Elhadad and Robin, 1996) is often considered high. In fact, the de-

[1] Of course, these texts are defined by the application, viz. the customer, as all other output.

```
[(SENTENCE DECL)
 (VC [(VOICE PASSIV)
      (MOOD IND)
      (TENSE PRAESENS)
      (SBP S2)
      (STEM "verursach")])
 (DEEP-SUBJ [(TOP Y)
             (TY GENERIC-NP)
             (NUMBER SG)
             (DET INDEF)
             (NR V2)
             (GENDER MAS)
             (STEM "antagonismus")
             (PP-ATR [(LOCATIVE ...)
                      (GENDER NTR)
                      (STEM "Acetylcholin")
                      (DET WITHOUT)
                      (NUMBER SG)
                      (TY GENERIC-NP)
                      (PREP MIT)])
              (ADJ [(STEM "kompetitiv")
                    (POS ADJECTIVE)
                    (DEG POS)])])
 (DEEP-AKK-OBJ [(TY GENERIC-NP)
                (NUMBER PLUR)
                (DET DEF)
                (STEM "wirkung")
                (GENDER FEM)])]
```

Figure 4: TG/2 Input Expression Partly Corresponding to Figure 3. The material for “on the level of the muscarinic sights of these substances” would appear under DEEP-SUBJ.PP-ATR.LOCATIVE, but has been omitted for reasons of space. The representation contains content word stems and names for syntactic structures (SBP, NR features). Determiners and prepositions are also provided.

velopment from scratch of a shallow grammar for a small NLG application on the basis of a simple framework like TG/2 can be more cost-effective.

Shallow and in-depth generation tasks can be related with help of TG/2. As the amount of domain-specific canned text in the TG/2 grammars correlates to the shallowness of the input, the generation tasks described can be located on a scale that ranges from shallow to in-depth domain and input models. There are trivial systems at one end that just produce canned text according to triggers (e.g. system error reports). A bit further on the scale we find template-style systems, like the air quality report generator, which use canned text to make knowledge implicit in the input explicit. In-depth realizers with sophisticated grammars that do not use domain-specific canned text at all are located at the other end of the scale, such as the MUSI generator.

Why are shallow and in-depth interlinguas both viable? One obvious reason lies in the origin of the interlingua representations. Shallow representations usually originate from non-linguistic processing, such as accessing a database or interpreting some user interaction, whereas in-depth representations generally have a linguistic origin, e.g. from an NL parsing component.

More interestingly, the type of domain and application determines the depth of modeling. Air quality reports form a small and closed domain. Implicit knowledge is easy to make explicit. A shallow model, being inherently simple, is perfectly adequate. A complex functor-argument representation would mean a dramatic overshot for this type of application. The same holds for many generation applications, such as reporting about stock exchange (Kukich, 1983) or weather forecasts (Boubeau et al., 1990). Medical scientific texts, on the other hand, form a very large domain, requiring broad-coverage linguistic knowledge. A shallow model would not even be able to capture the most frequent semantic relations. General means of expressing semantic relationships are mandatory.

What are the advantages and drawbacks of either approach? Shallow interlinguas allow for a straightforward multi-lingual generation. All linguistic processing can be concentrated in the module consuming the interlingua expression, e.g. TG/2. A drawback consists in domain-dependent grammars, which are hardly reusable for other applications. Still it is worthwhile, as the effort to create a grammar for another language is low.

With in-depth language-neutral representations, the issue of reusing existing linguistically motivated grammars arises, simply because of the tremendous effort for developing them from scratch. Technically an existing grammar may be reused if a well-defined interface is available. In TG/2, the interface to the input representations consists of the tests and access functions called from within the grammar rules. Depending on the different organization of information within input languages, this interface must be modified. If the same types of information required by the grammar can be produced by the new input language, the way is paved for a successful reuse. If the new input language offers different types of information, the adaptation problem described above arises.

4. On the Definition of Interlinguas

We now address issues on the semantics and pragmatics of interlinguas from a generation perspective by discussing three types of problems generators may encounter with in-depth interlinguas, using experiences with IRep4 as our source of examples.[2]

4.1. Extrinsic problems

In MUSI, a variety of problems with interlinguas known from machine translation were experienced, showing that this interlingua, as so many others, is not language-neutral in a strict sense. The problems were related to the fact that languages encode information differently and the interlingua cannot sufficiently abstract away from this. More precisely, although IRep4 does not contain elements specific to any of the four languages involved, the analysis results reflected some grouping and nesting of phrases and clauses of the source language.

[2] By critically reviewing IRep4, we necessarily omit mentioning many excellent features that made it very useful for the challenging task of representing scientific text.

For instance, Italian (and English) uses post-nominal adjectival clauses that correspond to a post-nominal relative clause or pre-nominal adjectival modifiers in German (cf. Figure 5a). German does not have the possibility to linearize or nest several adjectival or participial clauses after the head noun. Moreover, large phrases in pre-nominal position are difficult to understand since the head noun is uttered only afterwards.

In IRep4, these clauses are typically represented as restrictive modifiers (RESTR), accompanied, in the case of a predicative concept, by the source-language specification CAT = ADJP. The generator follows the heuristic strategy of assigning small adjectival phrases to the pre-nominal adjective position and large ones to the post-nominal relative clause position. In the latter case, the CAT specification will be ignored, as a full sentence with a copula must be generated. A further requirement consists of the need for one argument of the adjective to be realizable as the relative pronoun.

The result is not satisfactory, as it can lead to recursive center-embedding causing bad readability (cf. Figure 5b). The sentence in Figure 5c is stylistically much better; it has fewer closing brackets in a sequence, which means less deep embedding and improved readability. Linguistically, it shows two extrapositions, i.e. the innermost relative clause (not bracketed further) occupies the post-field[3] of the embedding one, which in turn occupies the post-field of the main clause. The stylistically preferred solution would be to realize the innermost clause as a prenominal AP, while extraposing the larger clause as a relative clause, as in Figure 5d.

Another striking example of language differences experienced with IRep4 is the use of determiners. English text does not use always definite articles when they are mandatory in German. For instance, "features of malnutrition" should be translated into "Merkmale der Mangelernährung" (definite article included), whereas "features of chronic malnutrition" corresponds to "Merkmale chronischer Mangelernährung" (no article).

IRep4 does, of course, not represent definite articles when there are no such determiners in the source-language text. The generator uses as a general rule that "naked" generalized possessives – i.e. the head of a RESTRictive modifier that corresponds to a noun and does not have a determiner or a modifier – are automatically accompanied by a definite article, covering the above examples.

English "Treatment consisted in..." should translate to "Die Behandlung bestand aus...", using a definite article. In these cases, a decision within the generator on whether or not to use a definite article would rely on lexical semantic information about both the source and target language lexemes.

The obvious solution to the extrinsic problems is to complement the level of interlingua with a set of transfer rules specific for every pair of source and target language. This complicates the situation, but would, in MUSI, have led to considerable stylistic improvements of the generated sentences.

For shallow models, this problem simply does not exist.

4.2. Intrinsic problems

IRep4 also has a few intrinsic properties that affected generation. Most prominently, it does not represent scope and thematic, or constituent, order information. The scope of negation would be important for the proper placement of the negation particle. Moreover, the scope of modifiers is not represented. With the current, inherently flat representation, i.e. multiple modifiers at the same level of embedding, generation cannot decide between e.g. "the following clinical case" and "the clinical following case". Modifiers should be nested to express this information.

Deciding about word order in generation is relevant to represent the argumentative structure in complex sentences and ensure coherence. The order of constituents in the source language text is not marked in IRep4, which may cause a deviating target-language order in German. This can lead to a lack of textual coherence, if e.g. a modifier that starts the sentence appears at the end. Consider "upon objective investigation, the woman‘s face was red and congested", which was translated into "das Gesicht der Frau war rot und geschwollen bei objektiver Untersuchung", generating the introductory PP at the end. A possible subsequent anaphoric reference would be less felicitous than in the original text. In the absence of a super-ordinated text planning stage, interlingua expressions should specify thematic order, or constituent order, in the source language text.

German generation assumes a standard word order for active voice, unless other information is given. The standard word order does not take into consideration the complexity, or the "weight", of a constituent. A heavy-weight subject preceding a short object in a transitive sentence is often considered bad style. Based on heuristics about a constituent's "weight", passive voice could have been chosen within the generator, causing the short constituent to precede the complex one, which generally leads to more fluent text (cf. the example in Figure 3). An interlingua should include hooks to provide this information. IRep4 might indirectly allow a good estimate by counting concepts, arguments and modifiers; further investigation is needed to identify a reliable formula.

For shallow interlinguas, intrinsic problems of this kind do not exist, as they are entirely dealt with in the grammar.

4.3. Pragmatic problems

In this section, we sketch some issues that can take a lot of effort to create a shared understanding among the researchers looking at interlingua expressions from different perspectives.

A grammatically correct input sentence is a legitimate input to a parser. Few systems can deal with incorrect sentences in an error-tolerant way. For generation, in-depth interlingua expressions should be correct in a similar sense. A formal specification of the interlingua is required to define its syntax and, very importantly, its semantics. Generation requirements should be formally specified as well and

[3] The post-field follows the infinite verb complex in a German declarative sentence. This position can be occupied by one constituent.

a) [[In the clinical case described,] [the symptoms] [were] [caused] [by ingestion [of anticolinergic substances [probably contained [in the leaves [of plants [consumed a few hours before]]]]]]].

b) [[In dem beschriebenen klinischen Fall] [wurden] [die Symptome] [durch [Verzehr [von anticholinergen Substanzen, [[die] [die Blätter [der Pflanze], [die vor ein paar Stunden genossen wurden,] möglicherweise enthielten,]]]]] [verursacht]].

In the described clinical case were the symptoms by ingestion of anticolinergic substances, that-were in-the leaves of-the plants, that-were a few hours before consumed, possibly contained.

c) [[In dem beschriebenen klinischen Fall] [wurden] [die Symptome] [durch Verzehr [von anticholinergen Substanzen]] [verursacht], [[die] [die Blätter [der Pflanze]] möglicherweise enthielten, [die vor ein paar Stunden genossen wurden]]].

d) [[In dem beschriebenen klinischen Fall] [wurden] [die Symptome] [durch Verzehr [von anticholinergen Substanzen]] [verursacht], [[die] [die [vor ein paar Stunden genossenen] Blätter [der Pflanze]] möglicherweise enthielten]].

Figure 5: Stylistic Variations in Translation. Brackets indicate some syntactic structure. a) English original sentence; b) Corresponding sentence in German with APs realized as relative clauses, with inter-linear translation; c) Extraposition of the relative clauses beyond the respective verbs; d) Realization of the innermost clause as a prenominal AP.

should be part of the "pragmatics" of the interlingua. For instance,

- the omission of information about tense, aspect, determination and number may mean that a default applies;
- a personal pronoun must either refer to an antecedent, or be accompanied by information about gender, person and number;
- an expression realized as a relative clause must contain exactly one constituent with a plain coreference specification; this constituent will become the relative pronoun;
- etc.

During the development of IRep4, this effort was not spent due to shortage of resources.[4] While from an analysis viewpoint, some decent output looks more or less satisfactory, it is the details that make generation feasible or cause its failure. Most importantly, the interpretation of interlingua expressions in NLG should be functional. Different surface representations corresponding to the same interlingua expression should be considered as equivalent in meaning. If this fundamental principle is not maintained, translation is not guaranteed to be meaning-preserving.

An interlingua can support this principle by making meaning representation explicit. IRep4 unfortunately has a fairly abstract representation for PP adjuncts and modifiers. The scheme is "`Mod = [<name>, <Irep4-expression>]`", where `<name>` is taken from a finite set of strings that more or less denote the semantics of the modifier. These names can be interpreted unambiguously by generation, but analysis may encounter difficulties in relating prepositions and head nouns to them, if only little lexical semantic knowledge is available. In Figure 3, the same name `RESTR` is realized differently, depending on the part of speech used for the embedded concept. If it is a noun, the semantics is that of a generalized possessive, which is realized in post-nominal position in German. If it is an adjective, a prenominal adjectival modifier is usually generated. Other uses of `RESTR` were mentioned above. If two or more meanings are connected to one name, it may appear psychologically difficult to refrain from using this name as a waste-basket.

Pragmatic problems exist for shallow models as well, as shallow input expressions are partly produced by external systems. In the air quality report generator, measuring values are received as input from a database. Time series are occasionally shortened by aggregating information ("from 9.00 to 11.00: 6,7 μg/m^3"). During the development, we have not been aware of the systematic omission of certain half hour values in the database, which occasionally leads to awkward results: "at 9.00: 6,7 μg/m^3; at 9.30: 0 μg/m^3; at 10.00: 6,7 μg/m^3; at 10.30: 0 μg/m^3; at 11.00: 6,7 μg/m^3". We easily could have implemented another aggregation rule that leads to output like "from 9.00 to 11.00: 6,7 μg/m^3, with every half hour value at 0".

5. Conclusion

In this contribution, we have related multi-lingual to cross-lingual generation and discussed emerging problems for the definition of an interlingua. This discussion was based on experience gained from implementing NLG components for a multi-lingual report generator and a cross-lingual summarization system within the same framework, TG/2. Shallow interlinguas originate from non-linguistic processing. They usually carry implicit meaning that must be made explicit in the generation process. For relatively small-coverage, closed domains, such as air quality reports, weather reports, or stock market reports, it is adequate to write specialized grammars using domain-specific canned text for this purpose. In-depth interlinguas usually originate from linguistic analysis, as in machine translation. The nature of the interlingua is closely tied to the sophistication of

[4] It is debatable though whether the resulting difficulties have been resolved with less effort.

the generation task in hand.

While well-modularized generation systems can be easily adapted to shallow interlinguas, an in-depth interlingua is much more complex to work with, as so many distinctions need to be addressed. In this paper we have identified some NLG requirements on in-depth interlinguas. From the experience with the MUSI application, we have learned that it is worthwhile to formally specify NLG requirements on the interlingua at the outset.

For a new application involving multi-lingual or cross-lingual generation, the interlingua should be chosen, adapted or designed according to the kind of linguistic processing involved and in view of the depth of modeling envisaged. On the shallow/in-depth scale, it should be as shallow as possible.

6. References

John Bateman and Renate Henschel. 1999. From full generation to 'near-templates' without loosing generality. In (Becker and Busemann, 1999), pages 13–18. Also available at `http://www.dfki.de/service/NLG/KI99.html`.

John Bateman. 1997. KPML delvelopment environment: multilingual linguistic resource development and sentence generation. Report, German National Center for Information Technology (GMD), Institute for integrated publication and information systems (IPSI), Darmstadt, Germany, January. Release 1.1.

Tilman Becker and Stephan Busemann, editors. 1999. *May I Speak Freely? Between Templates and Free Choice in Natural Language Generation. Workshop at the 23rd German Annual Conference for Artificial Intelligence (KI '99). Proceedings*, Document D-99-01. Also available at `http://www.dfki.de/service/NLG/KI99.html`.

L. Boubeau, D. Carcagno, E. Goldberg, Richard Kittredge, and A. Polguére. 1990. Bilingual generation of weather forecasts in an operations environment. In *Proceedings of the 13 th International Conference on Computational Linguistics (COLING-90), Volume 1*, pages 90–92, Helsinki.

Stephan Busemann and Helmut Horacek. 1998. A flexible shallow approach to text generation. In Eduard Hovy, editor, *Nineth International Natural Language Generation Workshop. Proceedings*, pages 238–247, Niagara-on-the-Lake, Canada. Also available at `http://xxx.lanl.gov/abs/cs.CL/9812018`.

Stephan Busemann. 1996. Best-first surface realization. In Donia Scott, editor, *Eighth International Natural Language Generation Workshop. Proceedings*, pages 101–110, Herstmonceux, Univ. of Brighton, England. Also available at the Computation and Language Archive at `http://xxx.lanl.gov/abs/cmp-lg/9605010`.

Stephan Busemann. 2002. Language generation for cross-lingual document summarisation. In Huanye Sheng, editor, *International Workshop on Innovative Language Technology and Chinese Information Processing (ILT&CIP-2001), April 6-7, 2001, Shanghai, China*, Beijing, China, May. Science Press, Chinese Academy of Sciences.

Jo Calder, Roger Evans, Chris Mellish, and Mike Reape. 1999. ”free choice” and templates: how to geth both at the same time. In (Becker and Busemann, 1999), pages 19–24. Also available at `http://www.dfki.de/service/NLG/KI99.html`.

Michael Elhadad and Jacques Robin. 1996. An overview of SURGE: a reusable comprehensive syntactic realization component. In Donia Scott, editor, *Eighth International Natural Language Generation Workshop. Demonstrations and Posters*, pages 1–4, Herstmonceux, Univ. of Brighton, England.

Karen Kukich. 1983. Design and implementation of a knowledge-based report generator. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 145–150, Cambridge, MA.

Alessandro Lenci, Nuria Bel, F. Busa, Nicoletta Calzolari, E. Gola, M. Monachini, Alexandre. Ogonowsky, I. Peters, W. Peters, N. Ruimy, M. Villegas, and Antonio Zampolli. 2000. SIMPLE: a general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Alessandro Lenci, Ana Água, Roberto Bartolini, Stephan Busemann, Nicoletta Calzolari, Emmanuel Cartier, Karine Chevreau, and José Coch. 2002. Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project. In *Procs. Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, May.

Ehud Reiter and Chris Mellish. 1993. Optimizing the costs and benefits of natural language generation. In *Proc. 13th International Joint Conference on Artificial Intelligence*, pages 1164–1169, Chambery, France.

The UNL Distinctive Features: Inferences from a NL-UNL Enconverting Task

Ronaldo Teixeira Martins[*], **Lúcia Helena Machado Rino**[**],
Maria das Graças Volpe Nunes[***], **Osvaldo Novais Oliveira Jr.** [****]

[*]Núcleo Interinstitucional de Lingüística Computacional - NILC
Av. do Trabalhador São-Carlense, 400 - 13560-970 - São Carlos, SP, Brazil
ronaldo@nilc.icmc.sc.usp.br
[**]Departamento de Computação - Centro de Ciências Exatas e de Tecnologia - UFSCar
Rod. Washington Luiz, km 235 - Monjolinho - 13565-905 - São Carlos, SP, Brazil
lucia@dc.ufscar.br
[***]Instituto de Ciências Matemáticas e da Computação (ICMC) - Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400 - 13560-970 - São Carlos, SP, Brazil
mdgvnune@icmc.sc.usp.br
[****]Instituto de Física de São Carlos (IFSC) - Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400 - 13560-970 - São Carlos, SP, Brazil
chu@ifsc.sc.usp.br

Abstract

This paper reports on the distinctive features of the Universal Networking Language (UNL). We claim that although UNL expressions are supposed to be unambiguous, UNL itself is able to convey vagueness and indeterminacy, as it allows for flexibility in enconverting. The use of UNL as a pivot language in interlingua-based MT systems is also addressed.

1. Introduction

Machine Translation (MT) is one of the most controversial subjects in the field of natural language processing. Researchers and developers are often at odds on issues concerning MT systems approaches, methods, strategies, scope, and their potentialities. Dissent has not hindered, however, the establishment of tacit protocols and core beliefs in the area. It has often been claimed that[1]: 1) fully automatic high-quality translation of arbitrary texts is not a realistic goal for the near future; 2) the need of some human intervention in pre-edition of the input text or in post-edition of the output text is mandatory; 3) source language should be rather a sublanguage, and the input text should be domain- and genre-bounded, so that the MT system could cope with natural language ambiguity; 4) the transfer approach is more feasible than the interlingual one, since the latter, albeit more robust and economic, is committed to the somewhat insurmountable task of designing a perfect (universal) language, comprising any other one; 5) common sense and general knowledge on both the source and the target cultures are as important as linguistic information, like in Knowledge-Based Machine Translation Systems (Nirenburg et al., 1992); 6) existing human translations can be used as a prime source of information for the production of new ones, similarly to the Example-Based Machine Translation Systems (Furuse and Iida, 1992); 7) existing MT systems are not appropriate to monolingual users, although they can be used to facilitate, speed up or reduce the costs of human translation, or to produce quick and cheap rough translations that may help the users to get a very broad idea of the general subject of the text.

Many authors obviously do not endorse all the listed statements, specially the fourth one. Hozumi Tanaka (1993), for example, argues in favor of the interlingua-based approach, and so do the research and development groups involved in interlingua-based systems, such as ULTRA (Farwell and Wilks, 1993), KANT (Mitamura et al., 1993), or PIVOT (Okumura et al., 1993). These works, however, rather confirm the very general observation that commercially available MT systems (e.g., SYSTRAN, VERBMOBIL, DUET (Sharp), ATLAS I (Fujitsu), LMT (IBM), METAL (Siemens)) are primarily transfer-based.

The most serious arguments against the interlingua approach concerns its alleged universality and excessive abstractness (Hutchins and Somers 1992). In order to cope with multilinguality, the interlingua should put aside language-dependent structures (such as the phonological, morphological, syntactical and lexical ones) and work at the logical level, which is supposed to be shared by human beings. Even at such uppermost level, however, there seems to be cultural differences. Eco (1994) reports, for instance, the case for Aymara, a South-American Indian language which would have three truth values, instead of the two "normal" ones. Furthermore, it has been said that, even if one comes to find this kind of perfect language, it would be so abstract that it would not be cost-effective, since the tools for departing from natural language and arriving at the logical representation would be excessively complex.

In what follows, we present some extra evidence towards the feasibility of interlingua-based MT. The Universal Networking Language (hereafter, UNL), developed by Uchida et al. (1999), brings some distinctive features that may lead to overcome some of the bottlenecks frequently associated to the interlingua approach. Although UNL was not designed as an interlingua, and MT is only one of the possible uses for UNL, it has been claimed that multilingual MT systems can use UNL as a pivot language. In this paper, some of the distinctive features of UNL are analyzed. We build

[1] Most of these assumptions can be extracted from the Survey on the State of the Art in Human Language Technology (Cole et al., 1995). Of special interest are the articles concerning multilinguality by Martin Kay (8.1, 8.2) and Christian Boitet (8.3, 8.4).

upon the experience in developing the Brazilian Portuguese (hereafter, BP) UNL Server, a bilingual MT system for translating Portuguese into UNL and vice-versa.

This paper is organized as follows. Section 2 provides a brief introduction to the UNL approach and some of its premises. In Section 3 we describe an experiment in which human subjects were asked to enconvert sentences from Portuguese into UNL. Section 4 brings the general results of the experiment. One of them is specially addressed in Section 5. Some issues arising from the results are presented in Section 6. Conclusions are stated in Section 7. The reader is supposed to have previous information on the UNL Project and knowledge on UNL Specification (at http://wwww.unl.ias.unu.edu) is considered mandatory.

2. The Universal Networking Language

The Universal Networking Language (UNL) is "an electronic language for computers to express and exchange every kind of information" (Uchida et. al., 1999, p. 13). According to the UNL authors, information conveyed by each natural language (NL) sentence can be represented as a hyper-graph whose nodes represent concepts and whose arcs represent relations between concepts. These concepts (called Universal Words or simply UWs) can also be annotated by attributes to provide further information on the circumstances under which they are used.

In this context, UNL is not different from the other formal languages devised to represent NL sentence meaning. Its structure is said to suffice to express any of the many possible meanings conveyed by any sentence written in any NL. This does not mean, however, that it is able to represent, at the same time, all the possible meanings conveyed by the very same NL sentence. Instead, UNL is able to represent each of them independently, and it is by no means able to provide a single structure coping with all of them. In this sense, there will never be a single UNL expression that completely suffices the meaning correspondence to a NL sentence. Or else: no UNL expression will be ever completely equivalent to a NL sentence, since the latter, but not the former, will allow for ambiguity.

In the following section, we report on results of a BP-UNL enconverting task that has been carried out by BP native speakers. In this experiment, we observe evidences that BP sentences must be disambiguated in order to be represented as UNL expressions.

3. The Experiment

In August 2001, we carried out an experiment on BP-UNL enconverting that involved 31 BP native speakers, all of them graduate and postgraduate students. Most of them (over 95%) were Computer Sciences students, aging 21 to 42 years old (90% of them were under 30 years old).

The experiment was split into training (steps 1-4) and test sessions (step 5), as follows: 1) a very general description of the UNL structure; 2) a general presentation of the definitions provided for five relation labels by the UNL Specification (1999), namely, 'agt' (agent), 'cag' (co-agent), 'obj' (affected thing), 'cob' (affected co-thing), and 'ptn' (partner); 3) an individual exercise on the use of the presented relation labels, in which subjects were asked to identify 50 different relations appearing in different BP sentences, indicating the corresponding UNL relation labels; 4) a public discussion on the exercise results; and 5) a final individual test in which subjects were asked again to identify 30 different relations appearing in different BP sentences, through their correspondence with the very same set of UNL relation labels. In Step 3 and 5, the subjects had also the option of pinpointing the impossibility of identifying either a relationship or its corresponding relation label, by choosing a "catch all" alternative (see option (a) in Figure 1). This exercise aimed at providing the means for the subjects to understand and explore BP-UNL enconverting, concerning the relation labels identification. This was then reinforced in Step 4, which was supervised by a UNL specialist. As it can be observed, these steps aimed at Step 5, the actual BP-UNL assignment, focusing on specific relation labels. In this step, some of the BP sentences presented to the subjects in Step 3 have been replicated.

Altogether, this experiment has taken 1 hour and 40 minutes, considering a 20-minute interval between the training and test sessions. Steps 1 and 2 have last 20 minutes, and so has Step 3 alone. Step 4, the longest one, has taken 40 minutes. Step 5, the actual test, has taken another 20 minutes. The interval between training and test aimed at allowing for the subjects settling on UNL specification, since test has been totally unsupervised. This also justifies our replication of some of the BP sentences used in training.

An English version of the task proposed in Step 3 is presented in Figure 1 below.

Considering the information presented in the first part of this experiment, identify the following:
1) If the relation depicted between the words signaled in each of the sentences below belongs to the five-relation set discussed previously; and
2) If so, which relation label would most suitably describe the involved relationship.

Use, for reference, the following code:
a) if NO label describes the relationship between the signaled words;
b) if the label AGT (agent) is the most suitable one;
c) if the label CAG (co-agent) is the most suitable one;
d) if the label COB (affected co-thing) is the most suitable one;
e) if the label OBJ (affected thing) is the most suitable one;
f) if the label PTN (partner) is the most suitable one.

Figure 1. Instructions for identifying and classifying relations.

The 30-sentence set used in the test session, along with its corresponding English translation, is shown in Figure 2.

SENTENCES	
1.	A crise quebrou o empresário >> ???(quebrou, crise) *The crisis broke the business man.* >> ???(broke, crisis)
2.	A crise quebrou o empresário >> ???(quebrou, empresário) *The crisis broke the business man.* >> ???(broke, business man)
3.	A farsa acabou. >> ???(acabou, farsa) *The farce is over.* >> ???(is over, farce)
4.	A neve caía lentamente. >> ???(caiu, neve) *Snow felt slowly.* >> ???(felt, snow)

5.	Alugam-se casas. >> ???(alugar, casa) *Houses are rented* (also: *Someone rents houses*) >> ???(are rented, houses)
6.	Choveu canivete ontem. >> ???(choveu, canivete) *It rained knives yesterday >> ???(rained, knives) (Brazilian Idiom)*
7.	João jogou o vaso com Maria contra Pedro. >> ???(jogou, Maria) *John threw the bowl with Mary against Peter.* >> ???(threw, Mary)
8.	João jogou o vaso com Maria contra Pedro. >> ???(jogou, Pedro) *John threw the bowl with Mary against Peter.* >> ???(threw, Peter)
9.	João lutou com Maria para vencer a doença. >> ???(lutou,Maria) *John fought with Mary to win the disease.* >> ???(fought, Mary)
10.	João não teve filhos com Maria. >> ???(ter, João) *John did not have children with Mary.* >> ???(have, John)
11.	Maria esqueceu o dia do aniversário da filha. >> ???(esquecer, dia) *Mary forgot her daughter's birthday.* >> ???(forgot, birthday)
12.	Maria foi despedida. >> ???(despedir, Maria) *Mary was fired.* >> ???(fire, Mary)
13.	Maria lembrou Pedro do horário. >> ???(lembrou, horário) *Mary remembered Peter about the schedule.* >> ???(remembered, schedule)
14.	Maria morreu com a falta de oxigênio.. >> ???(morreu, falta) *Mary died with the lack of oxygen.* >> ???(died, lack)
15.	Maria namorou Pedro. >> ???(namorou, Maria) *Mary flirted (with) Peter.* >> ???(flirted, Mary)
16.	Maria não foi ao cinema com a vizinha. >> ???(foi, vizinha) *Mary did not go to the cinema with her neighbor.* >> ???(go, neighbor)
17.	Maria não quis matar Pedro! >> ???(matar, Maria) *Mary did not intend to kill Peter.* >> ???(kill, Mary)
18.	Maria não se sentiu bem. >> ???(sentir, Maria) *Mary did not feel well.* >> ???(feel, Mary)
19.	Maria nunca conquistou Pedro. >> ???(conquistou, Pedro) *Mary never conquered Peter.* >> ???(conquered, Peter)
20.	Maria parece cansada. >> ???(parece, Maria) *Mary looks tired.* >> ???(looks, Mary)
21.	Maria se esqueceu de João. >> ???(esquecer, João) *Mary forgot John.* >> ??(forgot, John)
22.	Maria se matou. >> ???(matou, Maria) *Mary killed herself.* >> ???(kill, Mary)
23.	O filme deu origem a muitas controvérsias. >> ???(deu, filme) *The movie raised many controversies* >> ???(raised, movie)
24.	O frio congelou o pássaro. >> ???(congelar, frio) *The cold froze the bird.* >> ???(froze, cold)
25.	O medo da morte provoca insônia. >> ???(provoca, medo) *Fear of death causes insomnia.* >> ???(causes, fear)
26.	O pai com os filhos matou a mãe. >> ???(matou, filhos) *The father with the children killed the mother.* >> ???(killed, children)
27.	O pássaro congelou com o frio. >> ???(congelar, frio) *The bird froze (i.e., was frozen) with the cold.* >> ???(froze, cold)
28.	Os carros se chocaram na estrada. >> ???(chocaram, carros) *The cars crashed each other on the road.* >> ???(crashed, cars)
29.	Pedro se parece com a mãe. >> ???(parece, mãe) *Peter looks like his mother.* >> ???(looks, mother)
30.	Precisa-se de funcionários. >> ???(precisar, funcionários) *Employees are needed.* (also: *Someone needs employees*) >> ???(need, employees)

* Students were presented only to the original Brazilian Portuguese sentence. In the translation from Portuguese into English we tried to preserve the Portuguese syntactic structure as often as possible, even when the resulting English sentence sounds agrammatical.

Figure 2. Test corpus.

4. Results

The results of the experiment were the following:

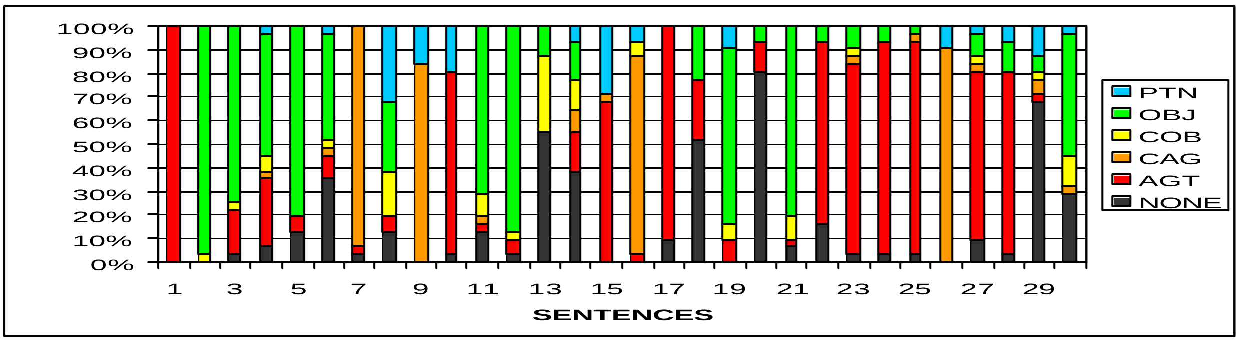


Figure 3. Distribution of BP-UNL enconvertings by subjects, with respect to the 5-relation labels set

Figure 4 below groups the results according to the agreement among enconverters.

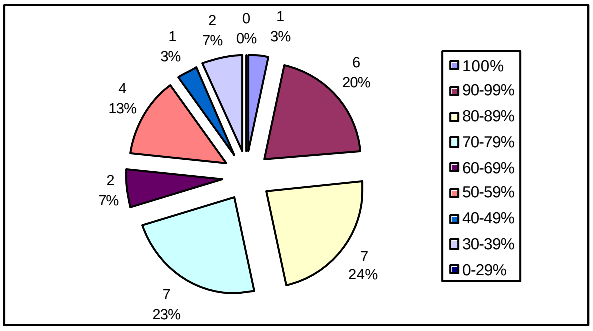


Figure 4. Agreement among enconverters.

A single relation (between "crise" (*crisis*) and "quebrou" (*to break*) in sentence 1: *"A crise quebrou o empresário" (= The crisis broke the business man*) led to an agreement of 100% among enconverters: they all used the 'agt' label in this case. There was an agreement between 90% to 99% on labeling relations in 6 sentences. Enconverters also agreed between 80% to 89% in assigning labels in 7 sentences. Other 7 sentences involved 70% to 79% agreement. In the remaining 9 sentences, agreement among enconverters was lower than 70%.

5. Case Study: Sentence 14

Sentence 14 ("Maria morreu com a falta de oxigênio." (literally: "Mary died with the lack of oxygen.") can be taken as a typical example of those involving considerable disagreement among enconverters. The relation between the verb "morreu" (to die) and the noun "falta" (lack) was encoded in varied ways, as follows: a) as an agent one (16%); b) as an object one (16%); c) as a co-object one (13%); d) as a co-agent one (10%); e) as a partner one (6%); and f) as none of the previous five relations (39%).

The unavoidable issue that follows from the above is why UNL labels were used in such apparently fuzzy way. Several reasons could be pinpointed here: a) the lack of expertise (or even of attention) of human enconverters', for they could not have had enough knowledge of language, or motivation, to carry on the experiment (although they are BP native speakers and seemed to be willingly helpful and interested in participating); b) the lack of clarity of the UNL Specification itself, even though there had been considerable discussion in the training session, for the problems posed by the enconverters to be tackled; c) the structure of the experiment itself, which was indeed too brief and too shallow to properly evaluate the human enconverters' performance; and, finally, d) the ambiguity of test sentences.

The analysis of the enconverters' choices certifies that disagreements are due to the latter point. Although it is unlikely for a BP speaker to say that 14 above, out of context, could have many different colliding meanings, the experiment has proved that apparently unambiguous sentences are unambiguous only apparently. Although eventually invisible, NL vagueness and indeterminacy would be pervasive in ordinary language,

Actually, none of the labels assigned to the relation between "morreu" (to die) and "falta" (lack) in sentence 14 could be considered wrong. The lack of oxygen could be understood in many distinct ways, such as:

a) an agent ("agt"), or the "initiator of the action" of "Mary dying" (or "killing Mary");

b) a co-agent ("cag"), or a "non-focused initiator of an implicit event that is done in parallel", in the sense it was not the lack of oxygen that killed Mary but either b.1) the situation (or the person) that has provoked the suppression of Mary's air supply or, in a more precise way, b.2) the reaction provoked (mainly in the brain) by the lack of oxygen;

c) an object ("obj") for the event described by "dying", since it is somehow "directly affected" by it, as the conclusion that the oxygen was lacking might be said to come directly from the fact that Mary died, otherwise no one would perceive that oxygen was lacking;

d) an affect co-thing ("cob"), or as being "directly affected by an implicit event done in parallel", if the observation that the oxygen was lacking were said not to come directly from the fact that Mary died, but from the fact that her lungs stopped working, which caused her to die;

e) a partner ("ptn"), for it could be somewhat "an indispensable non-focused initiator" of the action of "Mary dying", as if the main responsible for Mary's death was Mary herself (or someone else) that turned the oxygen suply off.

Besides such illustrations, many other relations can be said to hold between 'lack of oxygen' and 'die', namely, "met" (method), "man" (manner), "ins" (instrument), and "rsn" (reason), all easily applicable to such a case.

Such a variety proves that sentence 14 was indeed vague. The syntactic relation between the BP verb and its adjunct can convey many different semantic cases. Nevertheless, the UNL expression – whatever it may be – will have, in turn, a single interpretation, because relation labels are not supposed to overlap. The relations agt(die,lack), cag(die,lack), cob(die,lack), obj(die,lack), ptn(die,lack), although applicable to that very same NL sentence, are expected to label different (albeit related) phenomena. Indeed, to say agt(die,lack) is not the same as to say cag(die,lack) or ptn(die,lack). No intersection between these relations is envisaged in the UNL Specification, since they are meant to be exclusive[2].

This makes clear that the UNL specification forces filtering possible interpretations for NL sentences, in the sense a UNL expression must provide a completely unambiguous representation for the source sentence. As a matter of fact, although UNL is intended to be as expressive as any NL, UNL expressions cannot convey, at least at the relation level, NL vagueness and indeterminacy. Like any other formal language, UNL is committed to disambiguate NL sentences and, hence, to impoverish their semantic power.

Nevertheless, in no one of the above situations it is possible to say that a relation label is wrong, or that is completely inappropriate, although some of them may seem really unlikely to hold, depending on the context. The point is that the meaning of the sentence "Mary died with the lack of oxygen." is not encapsulated in the sentence itself but it is built out from the reading (and hence from the analysis) made by human enconverters. Since different enconverters have different underlying assumptions during their readings, the same BP phenomena can naturally imply different interpretations, which in turn lead to distinct UNL labeling. To conclude, it seems impossible to prevent subjectivity (or context-sensitiveness, or else, enconverter-sensitiveness) at that extent, no matter how univocal NL sentences seem to be.

6. Consequences

From the above it is possible to state that UNL should not seek for a straightforward correspondence between UNL expressions and NL sentences. It would be useless. As meaning is not encrypted in NL sentences but build through the analysis process, different enconverters will unavoidably propose different UNL expressions for the

[2] Accordingly, it is worthy to observe that the individuality of relations seems to be less strong when we consider other UNL relation labels set, e.g., that comprising "qua" (quantity), "nam" (name) and "pos" (possessor), which seems to be, to some extent and context, replaceable by "mod" (modification), implying that the latter can quite feasibly be at an uppermost level in a relation hierarchy. The same could be said of "met" (method) and "ins" (instrument), which seem to be under the scope of "man" (manner). Conversely, this does not mean that "mod" comprises any of "qua", "nam", or "pos", or that "man" embeds "met" and "ins". Instead, it does mean that both "mod" and "man" seem to share a comprehensive set of features with the relations that they replace. This is not the case of "agt", "cag", "cob", "obj", and "ptn", which seem to be in a more outstanding opposition.

very same NL sentence and many of these different expressions are legitimate.

Due to structure of UNL, UNL expressions cannot replicate NL sentence vagueness and indeterminacy. Enconverters are obliged therefore to choice a single interpretation among many different possible ones. This choice will be inevitably affected by the enconverters' context, which will be unreplicable itself by other enconverters. Once all these enconvertings will be valid, in the sense they are context-motivated, there will never be a one-to-one mapping between NL sentences and UNL expressions.

Accordingly, correctness, in UNL, instead of representing a (impossible) single possibility of enconverting, should rather be considered as fidelity to enconverters' intentions. UNL should clearly state that it would be up to the (human and machine) enconverter to decide what should the UNL representation be for a NL sentence. That is to say, the object of the UNL representation should be considered not exactly the meaning conveyed by the NL sentence but the *interpretation inferred by the enconverter from the use of that NL sentence in the enconverter's specific context*.

The fact that there could be more than a single (and adequate) UNL expression for the same NL sentence implies that UNL allows for flexibility in the enconverting process, although the UNL expression itself is not supposed to be flexible. It is up to the enconverter, and not the UNL specification itself, to decide which of the many possible interpretations is to be represented by a UNL expression. This is a significant UNL distinctive feature. Most formalisms do not allow for such variability and postulate that there should be a biunivocal relation between NL and its artificial representation. Otherwise, the formal representation would keep mirroring NL vagueness and indeterminacy, resulting useless.

The problem here is how to assure that enconverting flexibility will not prevent UNL from being a machine tractable language. As far as UNL expressions are dependent on the enconverter, there could be uncontrolled variations, which could blow out UNL into many different (and maybe mutually unintelligible) dialects.

This problem can be divided into two parts: 1) how to be sure that the UNL expression represents indeed what is intended by the enconverter; and 2) how to be able to generate, from such varied UNL expressions, NL grammatical sentences.

The first question is somewhat an educational problem. There are obviously misunderstandings and misuses of many relations. To say that it is up to the enconverter to decide which label should be used is not to say that the enconverter can do whatever he/she/it wants. The UNL Specification and other guidelines are to be followed. The relation "agt" must be applied to "a thing that initiates an action", and "ptn" should stand for "an indispensable non-focused initiator of an action". The relation "agt" cannot be used in a different sense: it would be wrong. Flexibility in encoding should not be mistaken for permissiveness. There are many correct UNL expressions for the same NL sentence, but there are also wrong UNL expressions.

The solution to such a problem cannot be, however, to state a rigid (a culture-, language-, context- and even enconverter-independent) relationship between a NL and UNL, otherwise UNL will not suffice to cope with inevitable varying enconvertings. The fact that meaning is build through the enconverting process and its main consequence, the fact that different enconverters will propose different expressions for the same NL sentence, should be both considered starting points, instead of something that one can or should avoid.

The best solution is, thus, to trust the enconverter (and maybe to certify enconverters), and to be conscious that, as in any other translation activity, there are good and bad translations, and bad translations do not prove that translating is not possible or that it does not work. Only time and enconverters' expertise can make UNL expressions better.

Nevertheless, to trust enconverters may imply making deconverting extremely difficult and costly. The more UNL allows flexibility in enconverting, the more costly will be UNL-NL deconverting, since the UNL expression may contain unexpected relations.

This is, however, a false problem. Deconverters are not committed to generate back the source sentence enconverted into UNL. Instead, they should be supposed to generate a NL sentence corresponding to the UNL expression. The original source sentence is definitely lost as it has been enconverted into UNL; only one of its possible interpretations (the one carried out by the enconverter) is preserved. Deconverters should take then UNL expression as the new source sentence, instead of using it just as an intermediate expression.

Furthermore, deconverting seems to be easier than enconverting, since much of the eventual meaning gaps may be inferred from the context by a human being (which is supposed to be the final user), instead of a machine. There is a very fragile break-even-point, from which generation results become excessively degraded, but the extent to which this happens will depend on the architecture of the UNL System.

7. Conclusion

The main conclusion to be extracted from the previous section seems to be a paradox: in multilingual MT Systems, in order to be a pivot language, UNL should not be treated as an interlingua, but as a source and a target language, at the same level as any other NL. Flexibility in enconverting brings UNL to be just like any other NL, in the sense it would allow UNL for coping with NL vagueness and indeterminacy, without sacrificing, however, the explicitness and clarity of UNL expressions, which would continue to be univocal and machine-tractable.

Acknowledgments

The authors acknowledge Mr. Tadao Takahashi, for his management of the Brazilian branch of the UNL Project, and to CNPq (Brazil) for the partial financial support. This work has also been partially supported by the UNU/IAS and the UNDL Foundation. The opinions expressed in the paper, however, do not necessarily represent the view of the UNL mentors.

References

Cole, R.A.; Mariani, J.; Uszkoreit, H.; Zaenen, A.; Zue, V. (Eds.) (1995). *Survey of the State of the Art in Human Language Technology*. NSF/CEC/CSLU. Oregon Graduate Institute. November.

(http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html)

Eco, U. (1994). *La recherche de la langue parfaite dans la culture européene*. Paris, France: Editions du Seuil.

Farwell, D. and Wilks, Y. (1993). ULTRA: A Multilingual Machine Translator. In S. Nirenburg (Ed.), *Progress in Machine Translation*. Washington, DC: IOS Press.

Furuse, O. and Iida, H. (1992). Cooperation between transfer and analysis in example-based framework. In *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, France.

Hutchins, W. J. & Somers, H. L. (1992). *An Introduction to Machine Translation*. San Diego, CA: Academic Press.

Mitamura, T., Nyberg, E. and Carbonell, J. (1993). In S. Nirenburg (Ed.), *Progress in Machine Translation*. Washington, DC: IOS Press.

Nirenburg, S., Carbonell, J, Tomita, M. and Goodman, K. (1992). *Machine Translation: A Knowledge-Based Approach*. San Mateo, CA: Morgan Kaufman.

Okumura, A., Muraki, K and Akamine, S. (1993). Multilingual Sentence Generation from PIVOT Interlingua. In S. Nirenburg (ed.), *Progress in Machine Translation*. Washington, DC: IOS Press.

Tanaka, H. (1993). Multilingual Machine Translation Systems in the Future. In S. Nirenburg (ed.), *Progress in Machine Translation*. Washington, DC: IOS Press.

Uchida, H., Zhu, M. and Della Senta, T. (1999). *Universal Networking Language: A gift for a millennium*. Tokyo, Japan: The United Nations University.

Structural and lexical transfer from an UNL graph to an equivalent natural language dependency tree

Etienne Blanc, Gilles Sérasset, WangJu Tsai

GETA, CLIPS-IMAG
BP 53, F-38041 Grenoble cedex 09
{etienne.blanc, gilles.serasset, tsai}@imag.fr

Abstract

We describe the transfer of an UNL graph into a equivalent tree, allowing to build UNL deconverters using existing MT systems based on tree processing.

1. Introduction

In the Universal Networking Language, a text is represented as a graph where nodes, bearing "Universal Words" (UWs), are linked by directed arcs bearing semantic "Relations Labels". A particular node, the "entry node", is distinguished in the graph.

The structure of these UNL graphs makes them quite suited to be processed by various linguistic tools. In particular, the Deconversion (from a UNL graph into an equivalent Natural Language text) or the Enconversion (from a Natural Language text into a UNL graph) may be achieved not only using the specially devised Deco and Enco tools, but also using adapted existing classical MT systems. For instance, UNL to Russian, UNL to Chinese, UNL to French deconverters are being developed using transfer MT systems.

Most of the classical MT systems use tree representation and not graph representation. Therefore the first step in the deconversion based on such systems is a graph-to-tree transfer. The aim of this paper is to discuss such a transfer, and to present the method used in the UNL-to-French deconverter.

We will begin by an overall presentation of the UNL-to-French deconvertor based on the ARIANE-G5 generator of MT systems. We will then discuss in more detail the process of graph-to-tree transfer.

2. A UNL-to-French deconverter deriving from a classical transfer system

2.1. Ariane-G5, a generator of MT systems

ARIANE-G5 is a generator of MT systems, that is an integrated environment designed to facilitate the development of MT systems (Boitet, 1997). These MT systems are written by a linguist using specialized languages for linguistic programming. ARIANE is not devoted to a particular linguistic theory. The only strong constraint is that the structure representing the unit of translation (sentence or paragraph) must be a decorated tree.

Fig.1 shows an overview of a classical transfer MT system using the ARIANE environment. The processing is performed through the three classical steps : analysis, transfer and generation.

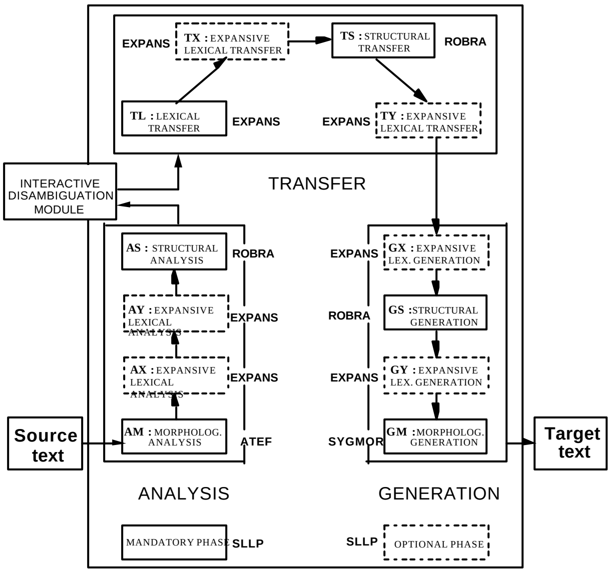

EXPANS
TX : EXPANSIVE LEXICAL TRANSFER
TS : STRUCTURAL TRANSFER
ROBRA
TL : LEXICAL TRANSFER
EXPANS
EXPANS
TY : EXPANSIVE LEXICAL TRANSFER
INTERACTIVE DISAMBIGUATION MODULE
TRANSFER
AS : STRUCTURAL ANALYSIS
ROBRA
AY : EXPANSIVE LEXICAL ANALYSIS
EXPANS
AX : EXPANSIVE LEXICAL ANALYSIS
EXPANS
AM : MORPHOLOG. ANALYSIS
ATEF
EXPANS
GX : EXPANSIVE LEX. GENERATION
ROBRA
GS : STRUCTURAL GENERATION
EXPANS
GY : EXPANSIVE LEX. GENERATION
SYGMOR
GM : MORPHOLOG. GENERATION
Source text
Target text
ANALYSIS
GENERATION
MANDATORY PHASE
SLLP
SLLP
OPTIONAL PHASE

Figure.1 The Ariane-G5 environment as used for generating a transfer MT SYSTEM

2.2. Principle of the French Deconverter

Fig 2 shows an overview of the UNL-to-French deconverter using the ARIANE environment.

The first step is a graph-to-tree transfer, achieving both:

- the graph-to-tree structural transfer necessary for the ulterior Ariane processing
- a lexical "Universal Words" to French words lexical transfer.

The resulting tree is a classical "deep tree" ready for generation.

This first structural and lexical step will be discussed in detail below. The following classic generation step will not be discussed here.

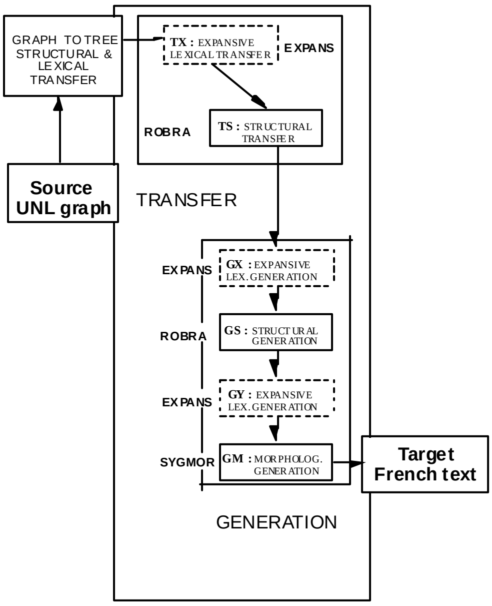

GRAPH TO TREE STRUCTURAL & LEXICAL TRANSFER
TX : EXPANSIVE LEXICAL TRANSFER
EXPANS
ROBRA
TS : STRUCTURAL TRANSFER
Source UNL graph
TRANSFER
EXPANS
GX : EXPANSIVE LEX.GENERATION
ROBRA
GS : STRUCTURAL GENERATION
EXPANS
GY : EXPANSIVE LEX.GENERATION
SYGMOR
GM : MORPHOLOG. GENERATION
Target French text
GENERATION

Figure 2 : The Ariane-G5 environment as used for generating a French deconverter.

3. UNL graph to NL tree structural transfer

The aim of the graph-to-tree structural transfer is to supply an output tree displaying all the structural information contained in the input UNL graph.

We will consider the following examples of tructural features encountered in a graph and needing some special coding in a tree are for instance:

- node having several mother nodes
- closed circuit
- hypergraph structure, that is graph containing nodes having themselves a graph structure (subgraphs, or "Compound Universal Words")

But before considering these examples, let's first illustrate the transfer on the simplest case, that is the transfer of a graph having in fact already a tree structure.

3.1. Graph with tree structure

In this simple case, the transfer is straightforward, as illustrated on figure 3.

This figure gives successively, from top to bottom:

- the meaning of the input graph as expressed in English
- the graph itself
- a sketch of its structure
- the structure of the equivalent tree as given by the structural transfer module (in this case the structure is the same as the structure of the graph)
- the decoration of the tree nodes.

The decoration of each node lists

- the Universal Word
- the semantic relation relative to its moither node (noted as a monovalued variable RSUNL)
- the attributes of the node (noted as a multivalued variable VARUNL)
- the id number (noted as the monovalued variable INST).

3.2. Graphs containing nodes with more than one mother node

In a tree, the root node has no mother node, and the other nodes have only one mother node. This is of course generally not the case for a graph, where all the nodes (including the entry one) may have several mother nodes.

Let's for instance consider the graph of fig. 4, where the entry node (« institute ») has a mother node (« establish ») the arc joining the first node to the second bearing the relation *obj*:

```
obj(establish(icl>found).@past,institute(ic
l>facilities).@present.@entry)
```

In order to get a tree, with a root node without mother node, the relation is inverted in the transfer module, and becomes

```
xxobj(institute(icl>facilities).@present.@e
ntry, establish(icl>found).@past)
```

where *xxobj* represents the inverse relation of the *obj* relation . The *obj* relation in the original graph expresses the fact that « institute » is the *obj* of establish, whereas the *xxobj* relation in the modofied graph expresses the fact that « establish » has « institute » as *obj.* Such an "inverted relation" is usally deconverted into French as a relative clause. The deconverted French text reads *"L'université des Nations Unie est un institut que l'Assemblée Générale des Nations Unies a fondé en 1975"."*

3.3. Graph containing a closed circuit

An equivalent tree structure of a graph containing a closed circuit may be obtained by opening the circuit, splitting one of its nodes as shown on fig.5 (the node "lecturer".splitted)

The new created node bears the same id number as the original one, indicating that it refers to the same object. In this example, this new node will be translated in French by the possessive "son", and the deconverter output reads *Le conférencier a lu son papier "*

3.4. Hypergraphs

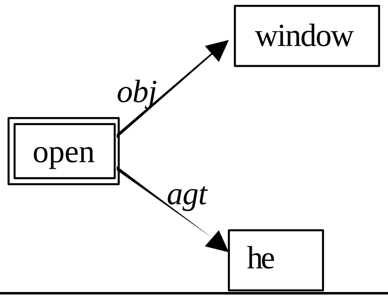
The processing of an hypergraph (graph containing subgraphs) is quite straightforward: the resulting tree is a tree containing subtrees.

English text: **He doesn’t open the window.**

Graph :

```
agt(open(icl>do).@entry.@not,he)
obj(open(icl>do).@entry.@not,window.@def)
```

Graph structure:



Output tree:

```
                                                  |-- 2:'WINDOW
                                                  |
                                      1:'OPEN' ---!-- 3:'HE'
```

Tree decoration:

```
1 'OPEN': VARUNL(ENTRY,NOT),INST(1)
    2 'WINDOW': VARUNL(DEF), RSUNL(OBJ),INST(1)
    3 'HE': RSUNL(AGT),INST(1)
```

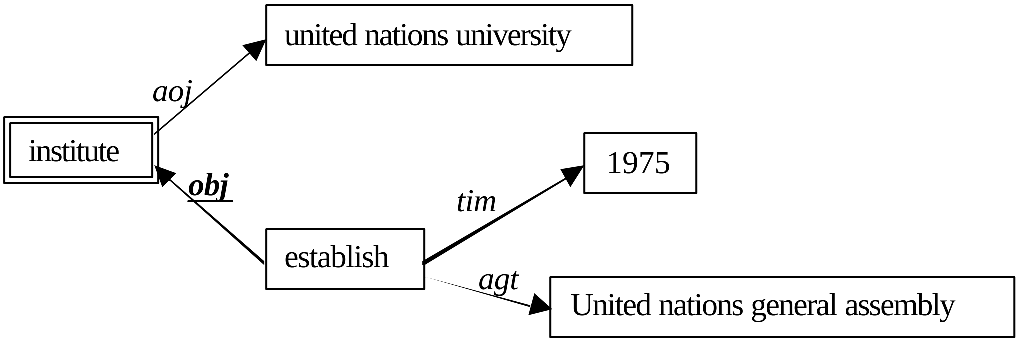
Fig 3. Structural transfer for a graph with tree structure..

English text: **The United Nations University is an institute which was founded by the United Nations General Assembly in 1975.**

Graph:

```
aoj(institute(icl>facilities).@present.@entry,united nations
university(icl>facilities))
obj(establish(icl>found).@past,institute(icl>facilities).@present.@entry)
agt(establish(icl>found).@past,united nations general
assembly(icl>organization))
tim(establish(icl>found).@past,1975)
```

Graph structure



Output tree

```
                       |-- 2:'UNITED NATION UNIVERSITY'
                       |                      |-- 4:'1975'
-- 1:'INSTITUTE' ------!-- 3:'ESTABLISH' -----!-- 5:'UNITED NATIONS GENERAL ASSEMBLY'
```

Tree decoration:

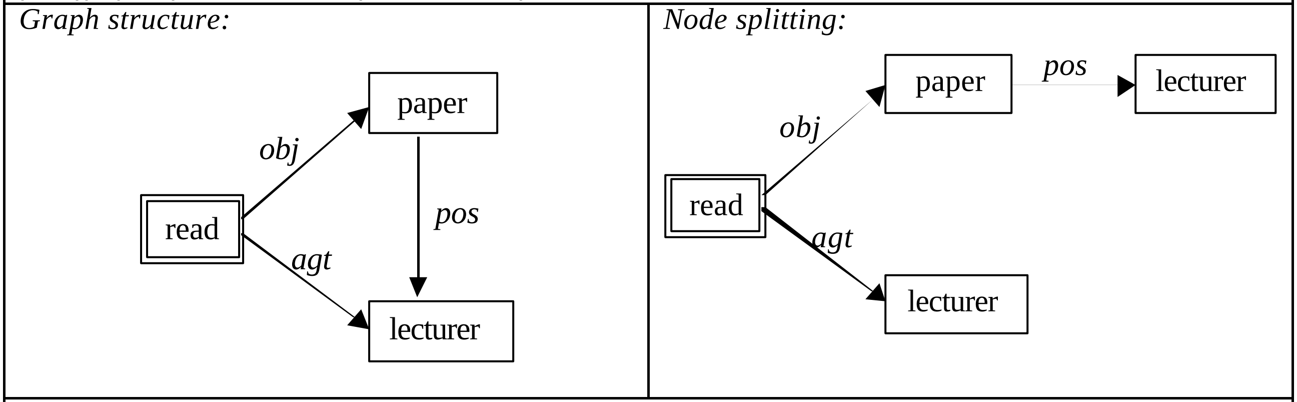
```
1 'INSTITUTE': VARUNL(PRESENT,ENTRY),INST(1)
    2 'UNITED NATIONS UNIVERSITY' RSUNL(AOJ),INST(1)
    3 'ESTABLISH': UL('<ESTABLISH>'), VARUNL(PAST), RSUNL(XXOBJ),INST(1)
      4 '1975':  RSUNL(TIM),CAT(CATCARD)
      5 UNITED NATIONS GENERAL ASSEMBLY': RSUNL(AGT),INST(1)
```

Figure 4 :Structural transfer of a graph whose entry node has a mother node

English text : **The lecturer read his paper.**

Graph :

```
agt(read(icl>do).@entry.@past,lecturer.@def)
obj(read(icl>do).@entry.@past,paper(icl>article))
pos(paper(icl>article),lecturer)
```



Graph structure:

Node splitting:

Output tree:

```
                          |-- 2:'PAPER' ----- 3:'LECTURER
                          |
            1:'READ' -----!-- 4:'LECTURER
```

Tree decoration

```
1 'READ': VARUNL(ENTRY,PAST),INST(1)
    2 'PAPER': RSUNL(OBJ),INST(1)
      3 'LECTURER': RSUNL(POS),INST(1)
    4 'LECTURER': VARUNL(DEF), RSUNL(AGT),INST(1)
```

Figure 5. Structural transfer for a graph containing a closed circuit.

4. UNL graph to NL tree lexical transfer

The structure of the UNL universal words makes in principle the lexical transfer a straightforward process.

A Universal Word like *mouse(icl>animal)* comprises indeed an headword "*mouse*" and a restriction "*icl>animal*" whose aim is to disambiguate the UW : distinction between *mouse(icl>animal)* and *mouse(icl>device).*

But in practice incompletness or inadequacies of the dictionaries leads either to use a treatment of the unknown word or an interactive lexical transfer.

4.1. Treatment of the unknown word

The treatment of the unknown words (that is of Uws whose NL language equivalents are not available in the dictionaries) may be based on the restriction of the UW and/or on the semantic relations the UW participates to.

4.1.1. Treatment of the unknown word based on the UW restriction

Using the restriction of the UW, we perform a partial treatment of the unknown word: the UW is not translated (the headword appears in the deconverted sentence), but the sentence is as far as possible correctly build.

This is shown on figure 6 where the graph contains two UWs supposed unknown. Testing the restrictions of the unknown UWs `rake(icl>do)` and `rake(icl>thing)` indicates that the first one is a verbal concept, the second one a thing concept, which allowed a correct construction of the sentence.

English text : **He rakes the leaves with the big rake.**

Graph :

```
agt(rake(icl>do).@entry,he)
obj(rake(icl>do).@entry,leaf(fld>bo
tany).@def.@pl)
ins(rake(icl>do).@entry,rake(icl>th
ing))
mod(rake(icl>thing),big(mod<thing))
```

French output text : **Il <<rake>> les feuilles avec le? grand? <<rake>>.**

Fig 6 Treatment of the unknown word based on the UW restrictions

4.1.2. Treatment of the unknown word based on the semantic relations

The semantic relations may also be used to determine the nature of the unknown word, allowing thus to obtain the correct sentence structure.

Figure 6 shows the deconversion result for a (unrealistic) graph where two unknown UWs without restrictions are present : *rake:01* and *rake:02* (the two different ids :01 and :02 indicate that these UWs are associated to two different nodes).

The different natures of both UWs were determined by using the semantic relations: the first instance of the UW rake, being the origin of an *agt* relation, was considered as a verbal concept, while the second one, being the target of an *ins* relation, was considered as a nominal concept.

English text **He rakes the leaves with the big rake.**

Graph :
```
agt(rake:01.@entry,he)
obj(rake:01.@entry,leaf(fld>botany)
.@def.@pl)
ins(rake:01.@entry,rake:02)
mod(rake:02,big(mod<thing))
```

French output text: **Il <<rake>> les feuilles avec le? grand? <<rake>>.**

Fig 6 Treatment of the unknown word based on the semantic relations.

4.2. Interactive lexical transfer

Our local deconverter may work in an interactive lexical mode. In this mode, for each UW in the graph, the French equivalent(s) present in the dictionaries are displayed for choice (figure 7).

```
Meeting(icl>event)
Click on one item below

Entering a new equivalent

meeting(icl>event)
réunion
CAT(CATN),GNR(FEM)

meeting(icl>event)
rencontre
CAT(CATN),GNR(FEM)
```

Figure 7 : Interactive lexical transfer

If no satisfactory equivalent is present in the dictionaries, the user may enter the correct equivalent, which is stored in an auxiliary dictionary, and becomes immediately available.

This interactive mode makes use of the PARAX-UNL hypertextual multilingual database (Blanc 1999)

5. Argument transfer

By argument transfer, we mean the relation between a UNL semantic relation and the corresponding syntactic function in the target natural language. It is not a one to one relation.

We will show here on an example how testing the restriction of a predicate may help finding the syntactic function associated to a semantic relation.

In the UNL language, one distinguishes the verbal concepts *do, occur, be.* For instance, the graph of fig. 8 contains the UW « open(icl>do) », whereas the graph of fig. 9 below contains the UW « open(icl>occur) ».

Both UWs are translated into French by the same verb, « ouvrir » (or in English by the same verb « to open »). But it is clear that in the case of « open(icl>do) », the subject syntactic relation for the French (or the English) verb corresponds to the *agt* relation (figure 8), but to the *obj* relation in the case of the « open(icl>occur) » UW.

That means that in such a case the restriction had to be tested in order to find the subject of the sentence.

He doesn't open the window.

```
agt(open(icl>do).@entry.@not,he)
obj(open(icl>do).@entry.@not,window
.@def)
```

Il n'ouvre pas la fenêtre.

Figure 8 The obj relation of this graph corresponds to the syntactic object relation in French or English

The window doesn't open.

```
[S]
;<SUZHOU_4>
obj(open(icl>occur).@entry.@not,win
dow.@def)
[/S]
```

La fenêtre n'ouvre pas.

Figure 9 The obj relation of this graph corresponds to the syntactic subjet relation in French or English

6. Conclusion

Such a UNL graph to Natural Language tree transfer proved to be quite feasible, and allowed us to reuse an existing French generator.

7. References

Boitet C. 1997 GETA's methodology and its current developments *PACLING'97, Meisei University, Ohme, Japan, sept 97, Proceedings 23-57.*

Blanc E (1999) PARAX-UNL, a large scale multilingual hypertextual database. *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium 1999 (NLPRS 99), pp 507-510. Tsinghua University Press, Beijing 1999.*

Some Lexical Issues of UNL

Igor Boguslavsky

Institute for Information Transmission Problems, Russian Academy of Sciences
19, Bolshoj Karetnyj, 101447, Moscow, Russia
bogus@iitp.ru

Abstract.

The Universal Networking Language (UNL) developed by Dr. H. Uchida at the Institute for Advanced Studies of the United Nations University is a meaning representation language designed for multi-lingual communication in electronic networks, information retrieval, summarization and other applications. We discuss several features of this language relevant for correct meaning representation and multi-lingual generation and make some proposals aiming at increasing its efficiency.

1. UNL approach to the lexicon.

The Universal Networking Language (UNL) developed by Dr. H. Uchida at the Institute for Advanced Studies of the United Nations University is a meaning representation language designed for multi-lingual communication in electronic networks, information retrieval, summarization and other applications.

Formally, a UNL expression is an oriented hypergraph that corresponds to a natural language sentence in the amount of information conveyed. The arcs of the graph are interpreted as semantic relations of the types agent, object, time, reason, etc. The nodes of the graph can be simple or compound. Simple nodes are special units, the so-called Universal Words (UWs) which denote a concept or a set of concepts. A compound node (hypernode) consists of several simple or compound nodes connected by semantic relations.

In addition to propositional content ("who did what to whom"), UNL expressions are intended to capture pragmatic information such as focus, reference, speaker's attitudes and intentions, speech acts, and other types of information. This information is rendered by means of attributes attached to the nodes.

After 6 years of the UNL project development, it is possible to take stock of what has been achieved and what remains to be done. In this presentation, I am going to concentrate on one of the central problems with which any artificial language is faced if it is designed to represent meaning across different natural languages. It is a problem of the language vocabulary.

I would like to single out three distinctive features of the UNL dictionary organization.

1. **Flexibility.** There is no fixed set of semantic units. There is only a basic semantic vocabulary that serves as a building material for free construction of derivative lexical units with the help of semantic restrictions. This makes it possible to balance to some extent the non-isomorphism of lexical meanings in different languages.

2. **Bottom-up approach.** The UNL dictionary consisting of Universal Words is not constructed a priori, top-down. Since it should contain lexical meanings specific to different languages, it grows in an inductive way. It receives contributions from all working languages. Due to this, one can expect that linguistic and cultural specificity of different languages will be represented more fully and more adequately than it would be possible under the top-down approach.

3. **Knowledge base.** As the UNL dictionary comprises unique semantic complexes lexicalized in different natural languages, we are facing the task of bridging the gap between them. It is supposed to be done by means of the Knowledge Base – a network of UNL lexical units connected by different semantic relations. Special navigation routines will be developed that will help to find the closest analogue to a lexical meaning not represented in the given language.

There are, however, some circumstances that impede full realization of these features, at least at the moment. Inductive storing of UWs from different languages is a good idea, but this process should be well organized. If a specific UW that is not self-evident is introduced to the UNL dictionary, it should necessarily be supplied at least by an informal comment to make it understandable to other users. Lucidity and easy interpretability of UWs is a goal at which all the developers of the UNL dictionary should aim.

Below, I am going to discuss in more detail two problems that have not so far received sufficient attention in UNL: the argument frames and lexical collocations.

2. Argument frames.

The need to introduce the information on the arguments does not seem to require justification. Any meaning representation language should have an ability to draw a distinction between the argument and non-argument links of predicates. In the UNL expressions, semantic links between the UWs are represented by means of UNL semantic relations. UNL disposes of an inventory of relations which, according to the latest specification, contains 41 items. Here are some examples of the UNL relations:

agt – agent (*John runs*),

obj – object (*read a book, A tree grows*),

ben – beneficiary (*He did not do anything for her*),

cag – co-agent (*I live with him*),

cob – co-object (*He fell into the river with the car*),

aoj – a thing which is in a certain state or is ascribed a property (*I love Mary; my brother is a student*).

dur – duration (*He worked nine hours*),

fmt – a range between two things (*He worked from Monday till Sunday*),

gol – final state (*turn red*),

ins – instrument (*observe with the telescope*),
met – method or means (*separate by cutting*),
pos – possession (*John's mother*),
rsn – reason (*They quarrel because of money*).

It is well known that for correct generation it is essential to know the argument structure of the predicates and the way each argument is expressed in the sentence. The UNL dictionary does not contain explicit information on the argument structure. According to the UW manual, the restrictions which should be included in the UW definitions are not meant for this purpose. As the UNL relations roughly correspond to semantic roles, it is supposed that each argument can be reliably identified based on its semantic role. However, this is not the case. Numerous attempts to construct a set of semantic relations, made over the last decades, showed that only a part of the relations between the words can be unambiguously interpreted in terms of semantic roles. In many cases this interpretation is largely arbitrary. This could not be a problem for the purposes of generation, if it were possible to assign semantic roles in a consistent way. Unfortunately, in practice it is hardly possible, especially when it is done by different people trained in different frameworks and working in different countries. The UNL texts compiled by the UNL project participants from 14 countries over the last years abound in mismatches in the representation of the same or very similar phenomena. Not surprisingly, most of them concern the representation of argument relations. For example, the phrase *base on respect* was interpreted by one team by means of the locative relation (lpl) and by another team by means of the comparative relation (bas), *freedom for all* was described with the purpose relation (pur) and with the beneficiary relation (ben), *bottleneck for the flow of information* received two labels – purpose (pur) and object (obj). Very often, the interpretation of a phrase in the corpus was motivated by the surface form rather than by its meaning. A typical example is *relations among nations* which was described by means of the locative relation obviously under the influence of the literal meaning of *among*. However, nations are by no means the place where relations occur. Rather, nations are participants of the "relations" situation and therefore are more likely to be objects (obj).

Sometimes the motivation behind the use of certain relations may be difficult to understand (at least, this is the case for the author of this paper). For example, in one of the sentences of the corpus, the argument structure of the verb *prevent* was presented as follows:

(1) *Nothing* (obj) *prevents members* (ben) *from discussing* (gol) *this problem*.

In our opinion, these problems are rooted not so much in the erroneous use of relations as in the fundamental impossibility of a consistent interpretation of all argument relations in terms of a small number of semantic roles.

What could one do to avoid the mismatches?

First, one could renounce using semantic roles in cases in which they are not obvious and replace them by semantically uninterpreted relations (subject, first object, second object, etc.). In this case, sentence (1) will receive a more transparent representation:

(2) *Nothing* (subject) *prevents members* (1 object) *from discussing* (2 object) *this problem.*

Obviously, it will be in many cases easier for those who write UNL expressions to develop a common approach to deciding which argument is the first object and which is the second than a common approach to finding appropriate semantic roles for them.

Second, one could accept the proposal of the French team and assign special markers to the case relations when they attach arguments (for example, @A would correspond to the first argument, @B – to the second, etc.). In this case, sentence (1) would be represented as:

(3) *Nothing* (obj.@A) *prevents members* (ben.@B) *from discussing* (gol.@C) *this problem.*

This would certainly reduce the area of uncertainty, but not eliminate it completely. To be able to interpret representation (3), the deconverter should know in advance the argument frame of the UW *prevent*. Otherwise, the uniformity of interpretation will still not be ensured. The only way to eradicate any ground for discordance between different users of the UNL language is to LIST ALL THE ARGUMENT STRUCTURES IN THE UNL DICTIONARY.

To incorporate this proposal, one need not introduce to the dictionary format any new possibilities: the existing apparatus of restrictions is quite sufficient. The only – but very serious – problem is to acknowledge that the argument frame should be explicitly and systematically specified in the UWs. If this is done, then one could keep using semantic roles in all the cases. For example, the word *bottleneck* (in the meaning of an obstacle) can receive the information that its syntactic object (*for something*) has the semantic role "pur" (or any other role which seems appropriate to the lexicographer). If every predicate is supplied with this information in the UNL dictionary, the discordance of opinion between different UNL users will become their private concern and the uniform treatment of the UNL relations in the most controversial zone – that of the argument relations – will be fully assured.

It should be emphasized however that in a general case the marking of the argument frame in a UW is not sufficient either. In some cases the same relation can attach to a UW both an argument and a free adjunct. For example, emotional states (of the type *be afraid, be surprised, be angry,* etc.) have an argument denoting a cause of the state. In sentence (4)

(4) *She is afraid to go out alone at night*

going out alone at night is what makes her to be in the state of fear. Therefore, relation "rsn" between *afraid* and *go out alone at night* is appropriate. On the other hand, *afraid* can have a non-argument cause, as in (5):

(5) *She is afraid (to go out alone at night), because this area is not very safe.*

Even if UW "afraid" is assigned a cause as one of the arguments (afraid(rsn>*)), we should know whether or not a "rsn"-link in the UNL expression denotes this argument. A good solution would be to mark the argument relation by a special label, as proposed in (3). Then, (5) will be represented as (6):

(6) rsn.@A(afraid(rsn>*), go out)
rsn(afraid(rsn>*), safe)

3. Lexical collocations.

Lexical collocations pose a serious problem for any language designed for representing meaning. Here are some examples of collocations from English: *give a lecture, come to an agreement, make an impression, set a record, inflict a wound; reject an appeal, lift a blockade, break a code, override a veto; strong tea, weak tea, warm regards, crushing defeat; deeply absorbed, strictly accurate, closely acquainted, sound asleep; affect deeply, anchor firmly, appreciate sincerely.* For simplicity, I will only dwell below on verbal collocations.

One of the problems such collocations raise is as follows. Some of the members of these collocations do not have a full-fledged meaning of their own. For example, the verb *give* in the collocation *give a lecture* does not denote any particular action. Its meaning, or rather its function, is the same as that of *take* in the collocation *take action*, or that of *make* in *make an impression*. The verbs *give, take* and *make* in these collocations are practically completely devoid of any meaning. Still, they have a very definite function – that of a support verb. This function is exactly the same in all the three cases, and nevertheless the verbs are by no means interchangeable. One cannot say **take an impression, *give action* or **make a lecture.* Moreover, this function is not only performed by different verbs with respect to different nouns. Very often, similar nouns in different languages require different verbs. For example, in Russian a lecture is not given but read, an action is not taken but accomplished, an impression is not made but executed.

How should these phenomena be treated in UNL? In particular, what UWs should be used for support verbs? The current practice suggests that UWs should be constructed on the basis of the source languages. Each language center should produce UWs for the words of its language, without any regard to other languages or any general considerations. A UNL expression and the UWs it consists of are considered adequate if they allow generating a satisfactory text in the same language they originated from. To what extent is this approach applicable to lexical collocations?

To answer this question, we will consider a concrete example. Suppose we have to convert to UNL Russian sentences with the meaning (7), (8), (9) or (10):

(7) *They began the war.*

(8) *We began the battle.*

(9) *The army suffered heavy losses.*

(10) *He took a shower.*

The problem is that in these contexts Russian uses quite different verbs than English. In Russian, correct sentences would be:

(7a) *They undid (razvjazali) the war.*

(8a) *We tied up (zavjazali) the battle.*

(9a) *The army carried (ponesla) heavy losses.*

(10a) *He received (prinjal) a shower.*

If UWs for support verbs in sentences (7a) – (10a) are constructed on the basis of Russian, they would look as follows: "undo(obj>war)", "tie up(obj>battle)", "carry(obj>loss)", and "receive(obj>shower)". These UWs will allow the Russian deconverter to produce perfect Russian sentences (7a) - (10a). In this case, the condition for adequacy mentioned above is met. Still, I would not consider UNL expressions based on these UWs adequate. They are produced without any regard for anything except the needs of Russian deconversion and are not fit for other purposes. In particular, these UWs are incomprehensible for anybody except Russians and it is doubtful that any other deconverter will be able to produce acceptable results from them. UWs originating from English will probably look like "take(obj>shower)", "begin(obj>thing)", "suffer(obj>loss)". To generate English sentences (7) - (10) from the UNL expressions constructed on the basis of (7a) – (10a), one would need to somehow ensure the equivalence of UWs "carry(obj>loss)" and "suffer(obj>loss)" in the Knowledge Base. This does not seem to be a natural and easy thing to do. Therefore, UWs for support verbs should not be constructed based on the lexical items of the source language.

Another possibility would be to make use of the co-occurrence properties of English lexical items. UNL vocabulary employs English words as labels for UWs and their meanings – as building blocks for UNL concepts which can be to a certain extent modified by means of restrictions. If lexical labels and meanings of UWs have been borrowed from English, their combinatorial properties can also be determined by the properties of corresponding English words. In this case, UWs and UNL expressions for sentences (7a) – (10a) will be identical to those for (7) – (10).

The advantage of this solution is obvious: since knowledge of English is indispensable for all the developers of X-to-UNL dictionaries, they can be sure that UWs for support verbs they produce are understandable and predictable. This solution has also drawbacks.

First, the inventories of support verbs in different languages are different. Therefore, we will often be faced with gaps in the lexical system of English and find no equivalent for a verb we need. Second, support verbs are bad candidates for the status of UWs. They do not denote any concept. Different support verbs often do not differ in meaning but only in their co-occurrence properties. It seems unreasonable to have different UWs to represent *take* (in *take action*), *make* (in *make an impression*) and *give* (in *give a lecture*), since the difference between these words is not semantic but only combinatorial. This difference should not be preserved in a meaning representation language.

The best solution would be to abstract from asemantic lexical peculiarities of support verbs and adopt a language-independent representation of these phenomena. Theoretical semantics and lexicography have long ago suggested a principled approach to the whole area of lexical collocations. It is the well-known theory of lexical functions by I. Mel'čuk implemented in the Explanatory combinatorial dictionaries of Russian and French (Mel'čuk 1974; Mel'čuk & Zholkovsky 1984; Mel'čuk et al. 1984, 1988, 1992, 1999). Possible use of lexical functions in NLP is discussed in (Apresjan et al. (in print)). Briefly, the idea of lexical functions is as follows. For more details, the reader is referred to the works mentioned above.

A prototypical lexical function (LF) is a general semantic relation R obtaining between the argument lexeme X (the keyword) and some other lexeme Y which is the value of R with regard to X (by a lexeme in this context we mean a word in one of its lexical meanings or some other lexical unit, such as a set expression). Sometimes Y is represented by a set of synonymous lexemes Y_1, Y_2, …, Y_n, all of them being the values of the given LF R with regard to X; e. g., MAGN (*desire*) = *strong / keen / intense / fervent / ardent / overwhelming.*

There are two types of LFs – paradigmatic (substitutes) and syntagmatic (collocates, or, in Mel'čuk's terms, parameters).

A substitute LF is a semantic relation R between X and Y such that Y may replace X in the given utterance without substantially changing its meaning, although some regular changes in the syntactic structure of the utterance may be required. Examples are such semantic relations as synonyms, antonyms, converse terms, various types of syntactic derivatives and the like.

A collocate LF is a semantic relation R between X and Y such that X and Y may form a syntactic collocation, with Y syntactically subordinating X or vice versa. R itself is a very general meaning which can be expressed by many different lexemes of the given language, the choice among them being determined not only by the nature of R, but also by the keyword with regard to which this general meaning is expressed. Typical examples of collocate LFs are such adjectival LFs as MAGN = 'a high degree of what is denoted by X', BON = 'good', VER = ‘such as should be’ and also support verbs of the OPER/FUNC family. Examples of the latter are OPER1 = ‘to do, experience or have that which is denoted by keyword X (a support verb which takes the first argument of X as its grammatical subject and X itself as the principal complement)’; OPER2 = ‘to undergo that which is denoted by keyword X (a support verb which takes the second argument of X as its grammatical subject and X itself as the principal complement)’; FUNC1 = ‘to originate from (a support verb which takes X as its grammatical subject and the first argument of X as the principal complement)’; FUNC2 = ‘to bear upon or concern (a support verb which takes X as its grammatical subject and the second argument of X as the principal complement)’.

If used in UNL, lexical functions will ensure a consistent, exhaustive and language-independent representation of support verbs and all other types of restricted lexical co-occurrence. For example, English and Russian support verbs we discussed above – *take* (*a decision, a shower*), *make* (*an impression*), *give* (*a lecture*), *suffer* (*losses*), *prinimat'* (*reshenie* ‘decision’, *dush* ‘shower’), *proizvodit'* (*vpechatlenie* ‘impression’), *chitat'* (*lekciju* ‘lecture’), *nesti* (*poteri* ‘losses’) – are correlates of the same lexical function – OPER1.

Being abstract and completely language-independent, lexical functions are devoid of all the drawbacks discussed above and can serve as an optimal solution to the problem of representation of the lexical collocations in UNL.

4. Acknowledgements.

This work has been supported by the Russian Foundation of Fundamental Research (grants Nos. 00-15-98866 and 01-07-90405).

5. References.

Apresjan Ju., I. Boguslavsky, L. Iomdin, L. Tsinman (in print). *Lexical function collocations in NLP.*

Mel’čuk I. A., 1974. *Opyt teorii lingvisticheskix modelej "Smysl – Tekst"* [A Theory of Meaning – Text Linguistic Models"]. Moscow, Nauka, 314 p.

Mel’čuk I. A., Zholkovskij A.K., 1984. *Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka.* [An Explanatory Combinatorial Dictionary of the Contemporary Russian Language] Wiener Slawistischer Almanach, Sonderband 14, 992 p.

Mel'čuk I., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Adèle Lessard, 1984. *Dictionnaire explicatif et combinatoire du français contemporain, Recherches lexico-sémantiques I.* Les Presses de l'Université de Montréal.

Mel'čuk I., Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, Suzanne Mantha, 1988. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II.* Les Presses de l'Université de Montréal.

Mel'čuk I., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha, 1992. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III.* Les Presses de l'Université de Montréal.

Mel’čuk I., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha et Alain Polguère, 1999. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV.* Montréal.

A rationale for using UNL as an Interlingua and more in various domains

Christian BOITET
GETA, CLIPS, IMAG
385, av. de la bibliothèque, BP 53
F-38041 Grenoble cedex 9, France
Christian.Boitet@imag.fr

LREC-02 First International Workshop on UNL, other Interlinguas and their Applications, 1 June 2002

Abstract

The UNL *language* of semantic graphs may be called as a "semantico-linguistic" interlingua. As a successor of the technically and commercially successful ATLAS-II and PIVOT interlinguas, its potential to support various kinds of text MT is certain, even if some improvements would be welcome, as always. It is also a strong candidate to be used in spoken dialogue translation systems when the utterances to be handled are not only task-oriented and of limited variety, but become more free and truly spontaneous. Finally, although it is not a true representation language such as KRL and its frame-based and logic-based successors, and although its associated "knowledge base" is not a true ontology, but rather a kind of immense thesaurus of (interlingual) sets of word senses, it seems particularly weel suited to the processing of multilingual information in natural language (information retrieval, abstracting, gisting, etc.).

The UNL *format* of multilingual documents aligned at the level of utterances is currenly embedded in html (call it UNL-html), and used by various tools such as the UNL viewer. By using a simple transformation, one obtains the UNL-xml format, and profit from all tools currently developed around XML. In this context, UNL may find another application in the localization of multilingual textual resources of software packages (messages, menu items, help files, and examples of use in multilingual dictionaries.)

Keywords: UNL, multilingual communication, cross-lingual information retrieval, localization

Introduction

UNL is the name of a project, of a meaning representation language, and of a format for "perfectly aligned" multilingual documents. There is some hefty controversy about the use of the UNL language as an "interlingua", be it for translation or for other applications such as cross-lingual information retrieval. On the other hand, there is almost no discussion on the UNL format, in its current form, embedded in HTML, or some directly derivable form, embedded in XML.

We argue that the UNL language is indeed a good interlingua for automated translation, ranging from fully automatic MT to interactive MT of several kinds through, we believe, spoken translation of non task-oriented dialogues. It is also more than that, due to the associated "knowledge base", and has a great potential in textual information processing applications.

We will first give our view of what the UNL language is, and then develop a "rationale" for using the UNL language UNL along the previous lines. We will then describe some interesting potential uses of the UNL format in an "XML-ized" form.

1. The UNL language

The UNL representation is made of "semantic graphs" where a graph expresses the meaning of some natural language utterance. Nodes contain lexical units and attributes, arcs bear semantic relations. Connex subgraphs may be defined as "scopes", so that a UNL graph may be a hypergraph.

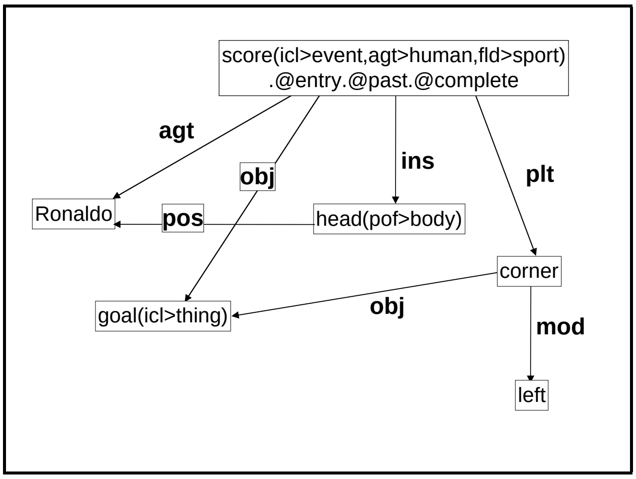


Fig. 1: a possible UNL graph for “Ronaldo has headed the ball into the left corner of the goal”

The lexical units, called Universal Words (in French, not "mot universel" but better "Unité de Vocabulaire Virtuel" or UVV or UW), represent word meanings, something less ambitious than concepts. Their denotations are built to be intuitively understood by developers knowing English, that is, by all developers in NLP. A UW is an English term or pseudo-term possibly completed by semantic restrictions.

A UW such as "process" represents all word meanings of that lemma, seen as citation form (verb or noun here). The UW "process(icl>do, agt>person)" covers the verbal meanings of processing, working on, etc.

The attributes are the (semantic) number, genre, time, aspect, modality, etc.

The 40 or so semantic relations are traditional "deep cases" such as agent, (deep) object, location, goal, time, etc.

One way of looking at a UNL graph corresponding to an utterance U-L in language L is to say that it represents the abstract structure of an equivalent English utterance U-E as "seen from L", meaning that semantic attributes not necessarily expressed in L may be absent (e.g., aspect coming from French, determination or number coming from Japanese, etc.).

2. Some arguments for using the UNL language in various contexts

To show that using UNL is not only a workable but a good or perhaps the best idea at the moment, we can say that

- the "pivot" technique HAS BEEN not only experimented but deployed successfully (ATLAS, PIVOT, ULTRA, KANT).
- in particular, ATLAS-II (Fujitsu) is built on the basis of a pivot from which the UNL representation has evolved. The main designer of UNL, H. Uchida, was also the main designer of ATLAS-II.
- ATLAS-II has been recognized as the best EJ/JE MT system in Japan for over 10 years and has a very large coverage (586,000 words in English and Japanese).
- interlingual representations can not in principle be used (alone) to achieve the highest quality achievable by transfer systems, BUT they can give quite high quality as demonstrated by ATLAS-II.
- due to the precise nature of UNL, it is possible for human non-specialists to improve a UNL representation interactively, a posteriori, from any UNL-related language, and on demand (meaning partially — think of "lazy improvement").
- in many contexts other than translation, an interlingual, semantic-oriented representation like UNL is actually the best solution. For example, all applications related to information processing in multilingual contexts don't need a very precise representation of the FORM of the information, they need a precise ENOUGH representation of the INFORMATION CONTENT of the information.

- applications such as information retrieval and abstracting have already been prototyped successfully with UNL. It is far easier to generate SQL or SQL-like queries and answers from a UNL form than from text in many languages.

3. Applications of the UNL format

The UNL *format* of multilingual documents aligned at the level of utterances is currenly embedded in html (call it UNL-html). A sentence is represented between the [S] and [/S] tags. Its original text is contained between {org:el} (English, here) and {/org}, its UNL graph between {unl} and {/unl}, each French version between {fr} and {/fr}, and analogously for other languages. Atrtibutes such as version, date, location, author, etc. may appear in the tags. Here is a slightly simplified example of a file in UNL-html format.

```
<HTML><HEAD><TITLE>
Example 1  El/UNL
</TITLE></HEAD><BODY>
[D:dn=Mar Example 1, on= UNL French,
mid=First.Author@here.com]
[P]
[S:1]
{org:el}I ran in the park yesterday.{/org}
{unl}
agt(run(icl>do).@entry.@past,i(icl>person))
plc(run(icl>do).@entry.@past,park(icl>place).@def)
tim(run(icl>do).@entry.@past,yesterday)
{/unl}
{cn dtime=20020130-2030, deco=man}
我昨天在公園裡跑步                    {/cn}
{de dtime=20020130-2035, deco=man}
Ich lief gestern im Park. {/de}
{es dtime=20020130-2031, deco=UNL-SP}
Yo corri ayer en el parque.{/es}
{fr dtime=20020131-0805, deco=UNL-FR}
J'ai couru dans le parc hier. {/fr}[/S]
[S:2]
{org:el}My dog barked at me.{/org}
{unl}
agt(bark(icl>do).@entry.@past,dog(icl>animal))
gol(bark(icl>do).@entry.@past,i(icl>person))
pos(dog(icl>animal),i(icl>person))
{/unl}{de dtime=20020130-2036, deco=man}
Mein Hund bellte zu mir.{/de}
{fr dtime=20020131-0806, deco=UNL-FR}
Mon chien aboya pour moi. [/S] [/P][/D]
</BODY></HTML>
```

The French versions have been produced automatically while the German and Chinese versions have been translated manually.

The output of the UNL viewer for French is:

```
<HTML><HEAD><TITLE>
Example 1  El/UNL
</TITLE></HEAD><BODY>
J'ai couru dans le parc hier.
Mon chien aboya pour moi.
</BODY></HTML>
```

and will probably be displayed by a browser as:

Example 1 El/UNL

J'ai couru dans le parc hier. Mon chien aboya pour moi.

and similarly for all other languages.

The UNL viewer produces on demand as many html files as languages selected and sends them to any available browser.

The UNL-html format predates XML, hence the special tags like [S] and {unl}, but it is easy to derive from it an XML format and to transform the documents into an equivalent "UNL-xml" format. Then, using DOM and javaScript, it is possible to produce various views, including that of a classical viewer, a bilingual or multilingual editable presentation, and a revision interface where not only the text but the UNL graph and possibly other structures may be directly manipulated.

Let us take an example from an experiment performed for the "Forum Barcelona 2004" on documents in Spanish, Italian, Russian, French and Hindi. Hindi and Russian are not shown, but Japanese has been added by hand. The XML form is simplified.

```
<unl:S num="1">
<unl:org lg="cn">在博覽會之後，城市 將獲得一片海岸域 </unl:org>
<unl:unl>
<unl:arc> agt(retrieve(icl>do).@entry.@future, city) </unl:arc>
<unl:arc> tim(retrieve(icl>do).@entry.@future, after) </unl:arc>
<unl:arc> obj(after, Forum) </unl:arc>
<unl:arc> obj(retrieve(icl>do).@entry.@future, zone(icl>place).@indef) </unl:arc>
<unl:arc> mod(zone(icl>place).@indef, coastal) </unl:arc> </unl:unl>
<unl:cn> 在博覽會之後，城市 將獲得一片海岸域 </unl:cn>
<unl:el> After a Forum, a city will retrieve a coastal zone.</unl:el>
<unl:es> Ciudad recobrará una zona de costal después Foro. </unl:es>
<unl:fr> Une cité retrouvera une zone côtière après un forum. </unl:fr>
<unl:it> Città ricuperarà une zona costiera dopo Forum. </unl:it>
<unl:jp> フォーラムの後で, 都市は沿岸水域を取り出す。    </unl:jp>
</unl:S>
```

Correct sentences are produced by the deconverters from correct and complete UNL graphs.

Suppose for the sake of illustration that some UNL graph has been produced from a Chinese version, and does not contain definiteness and aspectual information. All results may be wrong wrt articles, and some wrt aspect.

The idea of "coedition" is applicable if there is a UNL graph associated with a segment one wants to modify. The goal is to share the revisions across languages, by reflecting them on the UNL graph, e.g.

- add ".@def" on the nodes containing "city", "Forum".
- replace "retrieve" by "recover" and add ".@complete" on the node containing it.

It is not possible in principle to deduce the modification on the graph from a modification on the text. For example, replacing "un" ("a") by "le" ("the") does not entail that the following noun is determined (.@def), because it can also be generic ("il aime la montagne" = "he likes mountains"). Hence, the technique envisaged is that:

- revision is not done by modifying directly the text, but by using a menu system,
- the menu items have a "language side" and a hidden "UNL side",
- when a menu item is chosen, only the graph is transformed, and the action to be done on the text is stored and shown next to its focus in the "To Do" zone,
- at any time, the new graph may be sent to the L0 deconverter and the result shown. If is is satisfactory, that shows that errors were due to the graph and not to the deconverter, and the graph may be sent to deconverters in other languages. Versions in some other languages known by the user may be displayed, so that improvement sharing is visible and encouraging.

New versions will be added with appropriate tags and attributes in the original multilingual document in UNL-xml format, or in a DBMS, so that nothing is ever lost, and cooperative working on a document is feasible. UNL may find another application in the localization of multilingual textual resources of software packages (messages, menu items, help files, and examples of use in multilingual dictionaries.)

Apart of the "coedition", there are many other portential applications of UNL, such as:

- crosslingual information retrieval, on which we are currently working,
- abstracting & gisting, which has been prototyped at NecTec and in India,
- localization of software packages: messages in multiple languages could be created from UNL graphs produced from a graphical interface or by enconversion, and then sent to appropriate deconverters.

For this last point, we have found how to represent messages including variables (such as integers, file names etc.), but not yet how to handle messages including morphological or even lexical variants (as "4 goda / 5 let" for "4 years / 5 years" in Russian).

Conclusion

The UNL language is an artificial interlingua, embeddable in html or xml formats for multilingual document representation and processing. Because of its both abstract and linguistic nature, the UNL language offers many more interesting potential applications than other types of interlingua such as task and/or domain specific interlingua.

The history of MT shows that UNL will also be usable in the context of high-quality MT, quality being obtained through typology specialization and/or interactive improvement, a priori (interactive disambiguation after all-path robust analysis) and/or a posteriori by coedition of the text in any language and the corresponding UNL graph.

References

Blanc É. & Guillaume P. (1997) *Developing MT lingware through Internet : ARIANE and the CASH interface*. Proc. of Pacific Association for Computational Linguistics 1997 Conference (PACLING'97), Ohme, Japon, 2-5 September 1997, 1/1, pp. 15-22.

Blanchon H. (1994) *Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mockup*. Proc. of 15th International Conference on Computational Linguistics, COLING-94, 5-9 Aug. 1994, 1/2, pp. 115—119.

Boitet C., Guillaume P. & Quézel-Ambrunaz M. (1982) *ARIANE-78, an integrated environment for automated translation and human revision*. Proc. of COLING-82, Prague, July 1982, North-Holland, Ling. series 47, pp. 19—27.

Boitet C. (1994) *Dialogue-Based MT and self-explaining documents as an alternative to MAHT and MT of controlled languages*. Proc. of Machine Translation 10 Years On, 11-14 Nov. 1994, Cranfield University Press, pp. 22.21—29.

Boitet C. & Blanchon H. (1994) *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup*. Machine Translation, Vol. 9, N° 2, pp. 99—132.

Boitet C. (1997) *GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects*. Proc. of PACLING-97, Ohme, 2-5 September 1997, Meisei University, pp. 23-57.

Boitet C., Réd. (1982) *"DSE-1"— Le point sur ARIANE-78 début 1982*. Contrat ADI/CAP-Sogeti/Champollion (3 vol.), GETA, Grenoble, janvier 1982, 616 p.

Brown R. D. (1989) *Augmentation*. (Machine Translation), Vol., N° 4, pp. 1299-1347.

Ducrot J.-M. (1982) *TITUS IV*. In *Information research in Europe. Proc. of the EURIM 5 conf. (Versailles),* edited by Taylor P. J., London, ASLIB.

Kay M. (1973) *The MIND system*. In *Courant Computer Science Symposium 8: Natural Language Processing,* edited by Rustin R., New York, Algorithmics Press, Inc., pp. 155-188.

Lafourcade M. (2001) *Lexical sorting and lexical transfer by conceptual vectors*. Proc. of MMA'01, 29-31/1/01, SigMatics & NII, Tokyo, 10 p.

Lafourcade M. & Prince V. (2001) *Synonymies et vecteurs conceptuels*. Proc.,, 29-31/1/01, SigMatics & NII, Tokyo, 10 p.

Maruyama H., Watanabe H. & Ogino S. (1990) *An Interactive Japanese Parser for Machine Translation*. Proc. of COLING-90, 20-25 août 1990, ACL, 2/3, pp. 257-262.

Melby A. K., Smith M. R. & Peterson J. (1980) *ITS : An Interactive Translation System*. Proc. of COLING-80, Tokyo, 30/9-4/10/80, pp. 424—429.

Moneimne W. (1989) *TAO vers l'arabe. Spécification d'une génération standard de l'arabe. Réalisation d'un prototype anglais-arabe à partir d'un analyseur existant*. Nouvelle thèse, UJF.

Nirenburg S. & al. (1989) *KBMT-89 Project Report*., Center for Machine Translation, Carnegie Mellon University, Pittsburg, April 1989.

Nyberg E. H. & Mitamura T. (1992) *The KANT system: Fast, Accurate, High-Quality Translation in Practical Domains*. Proc. of COLING-92, 23-28 July 92, ACL, 3/4, pp. 1069—1073.

Sérasset G. & Boitet C. (2000) *On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter*. Proc. of COLING-2000, Saarbrücken, 31/7—3/8/2000, ACL, 7 p.

Slocum J. (1984) *METAL: the LRC Machine Translation system*. In *Machine Translation today: the state of the art (Proc. third Lugano Tutorial, 2–7 April 1984),* edited by King M., Edinburgh University Press (1987).

Wehrli E. (1992) *The IPS System*. Proc. of COLING-92, 23-28 July 1992, 3/4, pp. 870-874.

A Platform for Experimenting UNL (Universal Networking Language)

Wang-Ju TSAI

GETA, CLIPS-IMAG
BP53, F-38041 Grenoble cedex 09 France
Wang-Ju.Tsai@imag.fr

Abstract

We introduce in this article an integrated environment, which provides the initiation, information, validation, experimentation, and research on UNL. This platform is based on a web site, which means any user can have access to it from anywhere. Also we propose an XML form of UNL document as the base of future implementation of UNL on the Internet.

1 Introduction

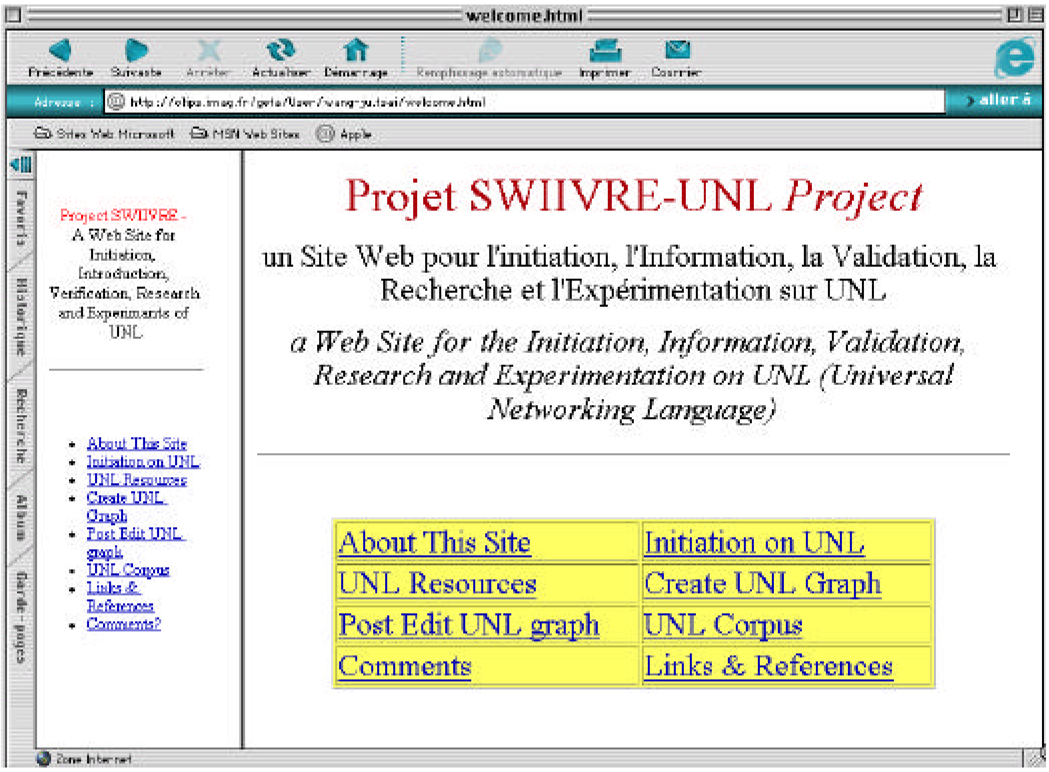
Since proposed 5 years ago, UNL project has attracted 16 international teams to join and is regarded as a very promising semantic Interlingua for knowledge representation on the Internet. The articles and applications of UNL have been found in many domain, such as: machine translation, information retrieval, multilingual document generation, ..etc. Now we can find on the Internet not only the web sites of UNL language centres but also some discussions. The applications to facilitate the usage of UNL have been produced as well. Now we see the need to create a platform to integrate these applications also to introduce UNL to new ordinary users. We create this platform on a web site SWIIVRE (http://www-clips.imag.fr/geta/User/wang-ju.tsai/welcome.html), which has several goals: for the initiation, information, verification, research, and experimentation of UNL. And since this platform is based on a web site, any user from anywhere can have access to it.

2 Introduction of the site SWIIVRE

In Appendix (I) we list all the resources accessible for UNL society members from internet. We can find out that most of the LC's connect vertically to UNL Centre but the horizontal connection among LC's is not enough, which means any user who wants to try the multilingualism of UNL will feel frustrated, since he will need to spend a lot of time try out every LC to know what service he can get.

The main purpose of this site is rather to integrate the current UNL applications and complete the services of Language Centres', when the function is available on a Language Centre, we simply provide the link to it, we also produce some applications to integrate or provide new functions, which all serve to facilitate the usage of UNL. Also we collect the useful information and publications on UNL, the web site is updated regularly. Lastly, by collecting the useful information and recording the related data, this site finally can serve as an evaluation of the performance of UNL community.

Here we show the welcome page of this site:



Project SWIIVRE - A Web Site for Initiation, Introduction, Verification, Research and Experiments of UNL

- About This Site
- Initiation on UNL
- UNL Resources
- Create UNL Graph
- Post Edit UNL graph
- UNL Corpus
- Links & References
- Comments?

Projet SWIIVRE-UNL *Project*

un Site Web pour l'initiation, l'Information, la Validation, la Recherche et l'Expérimentation sur UNL

a Web Site for the Initiation, Information, Validation, Research and Experimentation on UNL (Universal Networking Language)

About This Site	Initiation on UNL
UNL Resources	Create UNL Graph
Post Edit UNL graph	UNL Corpus
Comments	Links & References

The following is the introduction to each link on the welcome page:

2.1 About This Site:

This page provides the introduction, why and how this site exists, the site log and current status of this site, also the new projects to come on this site, lastly all the recent activities of UNL community. When clicked, a news flesh will also show the most recent UNL activities and the new updates on this site. In the future, we think we will at least UNL-ise this page to demonstrate the multilingualism of UNL.

2.2 Initiation on UNL:

This page is to help users to take a first step in UNL, understand how UNL works. We first provide a copy of most recent UNL specifications, for the moment only Spanish Centre has prepared a "multilingual interactive page" can serve as the tutorial and give examples to each UNL relations, thus we put a link to this page. When UNL becomes more well known, there will be more and more tutorials for beginners in the future. Or we might finally create an graphical interface for user to manipulate and show the spirit of UNL. We would also like to introduce the XML-UNL document here. We put an example of XML-UNL document here and with the help of XSLT, we can create the same effect like UNL browser, then the users can choose to read the document in the language they wish. We will explain later in the article why we want to XML-ise a UNL document.

2.3 UNL Resources:

This page provides all the UNL<->NL deconverters / enconverters, dictionaries that are accessible on the Internet. Some deconverters accept the deconversion of one single UW (Universal Word), in this case they can serve as the UNL-NL dictionaries. We can simply add some scripts in our site to help users to access these deconverters as if they are accessing dictionaries. In the future, the status report of each server will be added; we hope we can provide "UNL daily bulletin" to report the updates and status of each server. Currently only French server report can be seen. To complete the services, we developed a "multilingual simultaneous deconverter" (Preedarat 2001), which can handle several deconversions at one time. Users can click on the language versions they want as output, the program will contact these servers at once, thus they don't need to do the deconversions one by one, and they can experience the automatic multilingual generation.

2.4 Create UNL Graph:

Since ordinary users are not able to write UNL graph without being trained, to help users create UNL graph will be an important function to develop. In this page we collect the links to accessible UNL editors, including editor for professional writers or for beginners. We have put a link to our "Basic UNL graph editor" (Preedarat 2001), which is implemented by using a similar XML-UNL format and XSL transformation. The users can manipulate the UNL graph represented in tree-like structure, and save the result in XML format. We also put a link to the "interactive multilingual page" of Spanish Language Centre, here users can manipulate the UNL graph by the options provided, actually users can already generate many sentences based on these examples.

2.5 Post Edit UNL Graph:

This function is still under development. Our idea is to provide the users the possibility to correct the UNL document after it is deconverted. It provides ordinary users with the ability to correct the faults in the UNL graph and improve the quality of graph.

2.6 UNL corpus:

We collect all the UNL corpora here, and also we are currently working on designing a data base to store these corpora thus to facilitate the further exploitation or calculation. We can finally design an interface to allow users to upload the corpora in different forms, or produce the forms they desire. In appendix (II) is the first statistics we made on the corpus FB2004.

2.7 Comments:

To sends your comments to us.

2.8 Links & References:

We collect all the links to UNL Centre, Language Centres, articles, papers, discussion of UNL, and users can trigger the search engines here to find more information about UNL when they want.

3.XML-UNL document

The applications compatible to XML have been increasing a lot and XML can replace HTML as the next norm of a web-based document. And from an XML form, we can further produce other form, exchange or integrate the existing data easily. It would thus be reasonable to XML-ise the UNL document. We would like to propose here an XML form of UNL document as in Appendix (III). We created this DTD according to the UNL specification Version 3 Edition 1 (20/02/2002). Based on this DTD, we can create the UNL document in XML form, with an XSL Transformation we can produce the same effect as an UNL browser. Further more, we can easily expand this DTD to enable the XML-UNL document to register all the modifications and corrections on a UNL document, this can be very useful in our post-edition project.

Appendix (I)

4 Conclusion

We have made the first step in the integration of all the UNL components under a website. Next step is to streamline the procedures between current functions and to include more services.

5 References

Boitet Ch. (2001) Four technical and organizational keys for handling more languages and improving quality (on demand) in MT, " MT-SUMMIT VIII (2001) ", Proceedings of the Workshop (Towards a Road Map for MT), p.14-21. 18/09/2001

Coch & Chevreau (2001) Interactive Multilingual Generation. Proc. CICLing-2001 (Computational Linguistics and Intelligent Text Proceeding), Mexico, Springer, pp. 239-250.

Sérasset G. and Boitet Ch. (2000) "On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter", COLING 2000, Saarbruecken, Germany 31/07-04/08, p.768-774

Sérasset G. & BOITET Ch. (1999),"UNL-French deconversion as transfer & generation from an interlingua with possible quality enhancement through offline human interaction" MT Summit 99, 13-17 september 1999, Singapore, pp 220-228.

Boitet Ch. (1999) A research perspective on how to democratize machine translation and translation aids aiming at high quality final output, Machine Translation Summit VII (1999), Singapore, 13-17/9/99

Munpyo HONG & Olivier STREITER (1998) "Overcoming the Language Barriers in the Web: The UNL-Approach" , in 11.Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV'99), 1999, Frankfurt am Main.

Preedarat JITKUE (2001) Participation au projet SWIIVRE-UNL et première version d'un environnement web de déconversion multilingue et d'éditeur UNL de base, report de stage de Maîtrise Informatique Université Joseph Fourier – Grenoble 28/05-31/08

The resources accessible at each LC for UNL society members

	Enco	Deco	Dico	Introduction of UNL system	Linked by UNLC	remarks
Arabic	√	√	√	Arabic	√	
Chinese		√		English	√	
French		√				
Indonesian				Indonesian		
Italian		√		Italian	√	
Russian		√	√	English		
Spanish		√		English Spanish	√	Tutorials/Interactive Page/ Document Repository
Thai				Thai	√	
UNLC			√	English		UNL specs/ development modules

Appendix (II) Some Statistics about FB2004 Corpus

Corpus Name : FB2004
Original Language : English
Other available versions : French, Spanish, Italian, Russian, Hindi, UNL
No. of Sentences : 122
No. of Words : 2799
No. of Relations in UNL: 1519

Part I. The relation count

Relaion	**Outside scope**	**In scope**	**TOTAL**	**Relation**	**Outside Scope**	**In scope**	**TOTAL**
AGT	**66**	**10**	**76**	SEQ	0	0	0
AOJ	**64**	**37**	**101**	FMT	5	0	5
OBJ	**225**	**89**	**314**	FRM	6	3	9
AND	**63**	**120**	**183**	PLF	0	0	0
OR	26	3	29	SRC	2	0	2
BAS	2	2	4	GOL	17	7	24
CAG	0	0	0	PLT	1	0	1
CAO	0	0	0	TO	5	1	6
COB	1	1	2	INS	0	0	0
PTN	4	1	5	**MAN**	**49**	**17**	**66**
BEN	7	5	12	MET	10	3	13
PUR	28	1	29	PER	0	0	0
CNT	22	6	28	QUA	12	5	17
MOD	**263**	**186**	**439**	PLC	17	3	20
NAM	21	15	36	SCN	13	5	18
POF	5	2	7	TMF	2	0	2
POS	17	8	25	TMT	0	1	1
CON	2	0	2	VIA	1	0	1
RSN	1	0	1	DUR	5	4	9
COO	4	2	6	TIM	20	5	25

Total no. of relations							**1519**

Remarks:
a.)The 6 most frequently used relations are marked in bold type. The result is not surprising, since these relations have either an important or a broad usage. MAN and AGT's usage are frequent though straight forward. Besides its own static verb and copula usage, AOJ also shares part of adjective-noun relation, otherwise the frequency of MOD will be even higher.
b.)AND relation appears much more frequently within a scope, which is not surprising, since scope is used to represent the union of the similar things or ideas, and AND relation links these UW's in te scope.
c.)Some other relations' usage is not very braod, so they didn't appear.

Part II. Attribute count

(1)Time Attribute
.@past 40 / **.@present 114** / **.@future 187**
(2)Aspect Attribute
.@complete 20 / .@progress 13 / .@state 16 / else 0
(3)Reference Attribute
.@generic 9 / **.@def 659** / .@indef 79 / .@not 2 / .@ordinal 8
(4)Focus Attribute
.@entry 530 / .@topic 48 / .@title 21 / else 0
(5)Attitude Attribute
.@exclamation 1 / else 0
(6)Viewpoint Attribute
.@ability 7 / .@obligation 7 / .@possibility 8 / .@should 2 / .@unexpected-consequence 2 / else 0
(7)Convention Attribute
.@pl 558 / elso 0

Remarks:
a)The original text langue is English, so the frequency of .@pl, .@def, .@indef and time attributes are among the highest. If the original language is one of those isolated languages, such as Thai, Vietnamese, Chinese, which don't provide so much information about definitiveness or time, it might be difficult to use or to decide these attributes. It's not because that the graph authors or enconverters are bad, it's simply because they can't find these informations from the text when encoding.

Appendix (III)

```
<!DOCTYPE D [
<!ELEMENT D (P+) >
<!ELEMENT P (S+)>
<!ELEMENT S (org,unl,GS+)>
<!ELEMENT org (#CDATA)>
<!ELEMENT unl (#CDATA)>
<!ELEMENT GS (#CDATA)>

<!ATTLIST D dn CDATA #REQUIRED
            on CDATA #REQUIRED
            did CDATA #IMPLIED
            dt CDATA #IMPLIED
            mid CDTAT #IMPLIED>
<!ATTLIST P number CDATA #REQUIRED>
<!ATTLIST S number CDATA #REQUIRED>
<!ATTLIST org lang CDATA #REQUIRED
                    code CDATA #IMPLIED >
<!ATTLIST unl sn CDATA #IMPLIED
            pn CDATA #IMPLIED
            rel CDATA #IMPLIED
            dt CDATA #IMPLIED
            mid CDTAT #IMPLIED>
<!ATTLIST GS lang CDATA #REQUIRED
            code CDATA #IMPLIED
            sn CDATA #IMPLIED
            pn CDATA #IMPLIED
            rel CDATA #IMPLIED
            dt CDATA #IMPLIED
            mid CDTAT #IMPLIED>

]>

<!-- GS = generated sentence -->
<!-- dn = document name -->
<!-- on = owner name -->
<!-- did = document id -->
<!-- dt = date -->
<!-- mid = mail address -->
<!-- lang = lang tag -->
<!-- code = character code name -->
<!-- sn = system name -->
<!-- pn = post editor name -->
<!-- rel = reliability-->
]>
```

UCL – UNIVERSAL COMMUNICATION LANGUAGE

Carlos A. Estombelo Montesco
Dilvan de Abreu Moreira
Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação
Av. do Trabalhador São-Carlense, 400 - Centro - Cx. Postal 668 São Carlos - SP - Brazil CEP 13560-970
{cestombe, dilvan}@icmc.sc.usp.br

Abstract

For successful cooperation to occur between agents they have to be able to communicate among themselves. To enable this communication an Agent Communication Language (ACL) is required. Messages coded in an ACL should adequately express their meaning from a semantic point of view. The Universal Communication Language (UCL) can fulfill the role of an ACL and, at the same time, be convertible to and from a natural language. UCL design is concerned with the description of message structures, their underlining semantic context and the support for protocols for agent interaction. The key point about UCL is that the language can be used not only for communication among software agents but among humans too. This is possible because UCL is derived from the Universal Network Language (UNL), a language created to allow communication among people using different languages. UCL was defined using the Extended Markup Language (XML) to make it easier to integrate into the Internet. In addition, an enconverter-deconverter software prototype was written to serve as a tool for testing and experimenting with the language specifications.

INTRODUCTION

The technology of software agents can be an interesting tool for the creation of new models for complex software systems. In the project of software agents, many of the traditional techniques of artificial intelligence can be mixed with techniques from the field of distributed computer systems, theories about negotiation and theories about working teams (Dignum, 2000). Software agents are basically designed to cooperate (either with others or with humans) in a seemingly intelligent way. But for cooperation to occur a communication language is necessary.

What does it mean to be able to communicate with someone? Simplifying it, useful communication requires shared knowledge. While this includes knowledge of language, words and syntactic structures, meaningful communication is even more focused on knowledge about a problem to be solved. To interact with a florist you need some knowledge of flowers.

The widespread use of the Word Wide Web (WWW) and the growing Internet facilities has sparked enormous interest in improving the way people communicate using computers. To date, communication among software agents and humans has been done under limited conditions: communication is reduced to basic information exchange, ignoring the richness and flexibility implied by human language.

However to deal with any human language would be very difficult. To solve this problem, communication systems can use an Agent Communication Language (ACL) based in a simplified form of human language, which could be converted from and to a natural language.

OBJECTIVES

The main objective of this work is the specification of a new ACL, called UCL - Universal Communication Language, that focus on the specification of the semantic model and structure of the messages it represents. It also adds support for message transmission over the Internet and can be translated into or generated from natural language (English or other languages).

UCL is derived from the Universal Network Language (UNL) (Ushida et al., 1999) and implemented using the language XML (Extensible Markup Language) (Connolly, 2000). XML is a W3C (World Wide Web Consortium) standard language, like HTML, this means an easy integration with the Internet.

Another goal of this paper is to show a working UCL enconverter-deconverter prototype using the tool Thought Treasure and its associated ontology.

COMMUNICATION AMONG AGENTS

In the communication process among agents, it is indispensable an appropriate understanding of what will be communicated through the exchange of messages. A good representation of the knowledge domain, shared by the agents, can collaborate for a better understanding of the context where a message exchange takes place. As a consequence, it is important to explore concept classifications and their hierarchical structures for knowledge domain representation. The concepts in the knowledge domain have to be shared by the agents exchanging messages and be reusable in more than one context.

The specification of an ACL has to deal with the description of the message structure, his semantic model and the interaction protocols (Mamadou, 2000):

- The message format defines the communicative acts primitives and the parameters of the message (as sender, receiver, etc.). The message content describes facts, actions, or objects in a content language (KIF, Prolog, etc).

- The semantic model of an ACL should allow for messages with a concise meaning and no ambiguity.

- The interaction protocols are projected to facilitate the communication among agents. Protocols are optional, but, in case they are used, the communication among agents should be consistent with the chosen protocol.

ONTOLOGIES FOR COMMUNICATION

'Ontology' is a term used to refer to the common sense of some domain of interest. The ontology can be used as a uniform framework to solve communication problems.

An ontology necessarily links or includes some type of "general vision" regarding a certain domain. This "general vision" is frequently conceived as a group of concepts (for example: entities, attributes, processes), their definitions and their interrelations. That is called a conceptualization.

A conceptualization can be concretely implemented, for example, in a software component, or it can be abstract, being the implied concepts of a person. The use adopted in this work is that ontology is an explicit idea, or a representation (of some part) of a conceptualization.

An explicit ontology can take a variety of forms, but necessarily they will include a vocabulary of terms and some specification of their meanings (for example: definitions).

The level of formality for a vocabulary varies considerably. This variation can be shown in the following four points of view:

- Highly informal: expressed freely in natural language.
- Semi-informal: expressed in a restricted form and structure in natural language. Larger clarity for ambiguity reduction.
- Semi-formal: expressed in an artificial language defined formally.
- Strictly formal: defined meticulously with formal semantics, theorems and proofs.

A shared ontology is necessary for communication between two agents. Unfortunately UNL does not have a public available ontology. For this reason, the ontology embedded in the tool Thought Treasure was used to implement the enconverter-deconverter prototype.

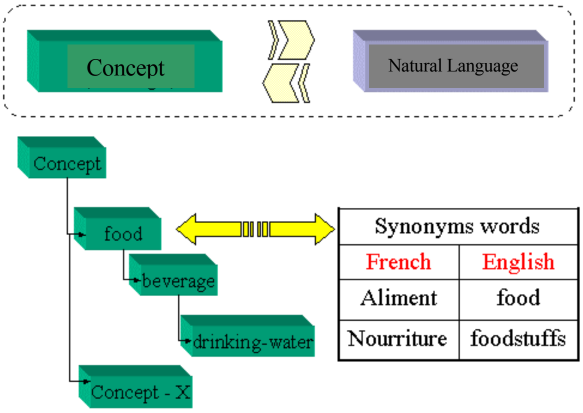
THE TOOL *THOUGHT TREASURE* (TT)

This is a powerful tool for processing natural language, developed by Erik T. Mueller (1998). It is capable of interpreting natural language, as well as extending its ontology-based knowledge base. TT has a compiler for natural language that allows it to extract information of sentences.

TT has a database with 25,000 concepts organized in a hierarchical way. For example, Evian is a flat-water type, which is a drinking-water type, which is a food type and so on.

Each concept has one or more word translations what forms a total of 55,000 words and sentences of the English and French language. For instance, as it is observed in the Figure 1, the association with the concept *food* in the English language are the words *food* and *foodstuffs* and in French *aliment* and *nourriture* (among others).

In addition, *ThoughtTreasure* has approximately 50,000 assertions related to concepts such as: a *green-pea* is a *seed-vegetable*, a *green-pea* is *green*, the *grean-pea* is part of *pod-of-peas*, and *pod-of-peas* is found usually at a store of foodstuffs.



Synonyms words	
French	English
Aliment	food
Nourriture	foodstuffs

Figure 1: Association of the ontology with a natural language

UCL - UNIVERSAL COMMUNICATION LANGUAGE

The language UCL represents information in the same way UNL does, but using syntax based in XML. XML is a meta-language, a simplified form of SGML, which developers can use to create new languages based in tag elements. The new tags, created to represent the new language elements, can be described in a special file called DTD (Document Type Definition). UNL is a formal language for representing the meaning of natural language sentences and exchange information over a network. Information that is written in a native natural language is "enconverted" into UNL and stored in a server. This information can be "deconverted" into other languages to be read by each native reader. Thus, UNL can play the role of an interface between different human languages to exchange information.

UNL represents information expressed in sentences as a set of relations between meanings, expressed by words, and a syntactic structure that makes up the sentence. The vocabulary of UNL consist of:

- Universal Words (UWs), to represent word meaning.
- Relation Labels, to represent relationships between UWs
- Attribute Labels, to express further definitions or additional information for the UWs that appear in a sentence.

In UNL, the information about a sentence includes its meaning, tense and aspect information (how the speaker grasp the event), intention of utterance, speaker's feeling or judgment upon contents, and sentence structure. In the language, the meaning of a sentence is represented by the description of the relationships between UWs and its structure is described by attaching attribute labels to these UWs.

UCL GOALS

The language UCL is to be used for high-level communication among agents through the exchange of messages. Some characteristics that guided the definition of the language were:

- To aid the communication involving agents giving importance to the semantics of the message;
- To be easy to use;
- To facilitate its integration into the Internet environment writing it in XML (*Extensible Markup Language*)

The language UCL represents the information in sentences (that can form messages) that involves a syntactic structure with a group of concepts, relationships and attributes similar to UNL:

- Universal Words (UW),
- Relationship labels,
- Attribute labels.

To define a language based in XML a specific DTD file is used. This DTD is essentially a grammar of free context, like the extended BNF form (*Backus Naur Form*) used to describe computer languages (Grosof & Labrou, 1999).

As in UNL, a Universal Word (UW) is the minimum unit that represents a concept, which denotes a specific meaning in a message. When a concept needs to be defined in more detail Relationship Labels and Attribute Labels are used. In addition, UCL uses a shared ontology, from the tool ThoughtTreasure, to add meaning to the UWs. All agents participating in a communication process should share this ontology.

In a UCL sentence, each defined UW has an identifier label (id) that is used to identify a particular concept inside a sentence. A sequence of alphanumeric characters forms this labels. The label `head` corresponds to the place where the name of the concept will be defined. The concepts used are always related to the ontology being used (ThoughtTreasure ontology). It is at this point that a sentence in UCL is connected with the ontology for a specify knowledge domain.

In UCL messages possess a certain meaning involving concepts. This composition of concepts is represented by groups of binary relationships, which allow different relationships involving the concepts. The relationship labels used come from UNL. Figure 2 shows an English sentence and its translation to UCL.

- *UNL is a common language that would be used for network communications.*

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE sentence SYSTEM "Sentence.dtd">
<sentence>
        <uw id="uw00" head=“language”>
            <icl direction=”to”>
                <uw head=”abstract thing”/>
            </icl>
            <tense attribute="present"/>
            <focus attribute="entry"/>
        </uw>
        <uw id="uw01" head=”UNL”>
            <icl direction=”to”>
                <uw head=”language”/>
            </icl>
            <focus attribute="topic"/>
        </uw>
        <uw id="uw02" head=”common”>
            <aoj direction=”to”>
                <uw head=”thing”/>
            </aoj>
        </uw>
        <uw id="uw03" head=”use”>
            <icl direction=”to”>
                <uw head=”do”/>
            </icl>
            <tense attribute="present"/>
        </uw>
        <uw id="uw04" head=”language”>
            < icl direction=”to”>
                <uw head=”abstract thing”/>
            </icl>
            <tense attribute="present"/>
            <focus attribute="entry"/>
        </uw>
        <uw id="uw05" head=”communication”>
            < icl direction=”to”>
                <uw head=”action”>
            </icl>
            <convention attribute="pl"/>
        </uw>
        <uw id="uw06" head=”network”>
            <icl direction=”to”>
                <uw head=”thing”>
            </icl>
        </uw>
    <relation label="aoj" uw-id1="uw00" uw-id2="uw01”/>
    <relation label="mod" uw-id1="uw00" uw-id2=”uw02”/>
    <relation label="obj" uw-id1="uw03" uw-id2=”uw04”/>
    <relation label="pur" uw-id1="uw03" uw-id2=”uw05”/>
    <relation label="mod" uw-id1="uw05" uw-id2=”uw06”/>
</sentence>
```

Figure 2 Definition a sentence in UCL

IMPLEMENTING AN ENCONVERTER-DECONVERTER

UCL is defined in the meta-language XML, to work with it a XML parser should be used. As the enconverter-deconverter is written in the language Java, the Java API for XML Processing (JAXP) Version 1.1 from Sun, was used (other Java XML parsers could have been used).

As said before, UCL uses the ontology available on the ThoughtTreasure (TT) tool (written in C). This tool includes program libraries to manipulate concepts of the ontology, to do consultations on the net of concepts, and to analyze their hierarchy. An instance of TT can run as a server in a network and communicate with a Java program running in another process. A Java communication API is supplied with TT to handle the low level details of this communication.

The enconverter-deconverter prototype uses the Java communication API to contact a running instance of TT and use its functionality. Those include natural language treatment, ontology queries, etc. A high level Java interface was written to communicate with the TT server (through the API) and implement the high level functions needed by the prototype. This interface is called *UclLanguage*.

Figure 3 presents a diagram with the sequence of events that happens when the prototype makes use of the interface *UclLanguage* to generate UCL messages.

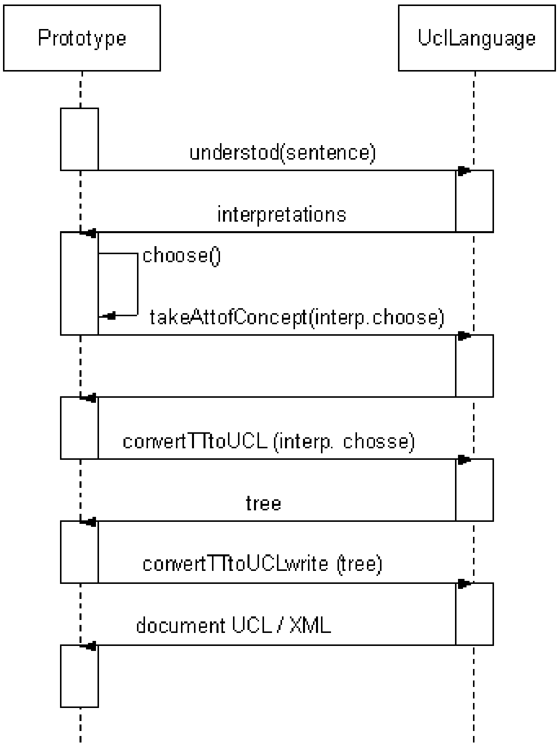

Prototype
UclLanguage
understod(sentence)
interpretations
choose()
takeAttofConcept(interp.choose)
convertTTtoUCL (interp. chosse)
tree
convertTTtoUCLwrite (tree)
document UCL / XML

Figure 3: Diagram with the sequence of events during enconvertion.

The process begins when a user enters a natural language sentence into the prototype. The prototype calls the method *understood* of the interface *UclLanguage*. The natural language sentence is interpreted (using TT) and some possible semantic interpretations are returned. The user chooses the most appropriate interpretation. The chosen interpretation is converted to TT's format (method *takeAttofConcept*) and then to UCL format (method *convertTTtoUCLwrite*). The UCL format can be shown on the screen or saved in a file.

The reverse process, transform a UCL message in natural language is easier. The prototype uses the method *deconvertUCLtoTT* to convert the UCL message in a list of TT concepts. Then it uses the method *deconverterTTtoLN* to transform this list of concepts in a natural language sentence, which represents the original UCL message.

Example : Monkey eats bananas

```
======= Input Natural Language ==========
Example: Monkey eats bananas.

============ Choose Option ==============
<0>An ape eats a banana.

Option: 0
============ Message UCL  ===============
<?xml version="1.0" encoding="UTF-8"?>

<sentence>
  <uw id="uw2" head="present-indicative">
    <icl direction="to">
      <uw head="present-tense" />
    </icl>
    <focus attribute="entry" />
  </uw>
  <uw id="uw4" head="eat">
    <icl direction="to">
      <uw head="ingest" />
    </icl>
  </uw>
  <uw id="uw5" head="ape">
    <icl direction="to">
      <uw head="mammal" />
    </icl>
  </uw>
  <uw id="uw7" head="banane">
    <icl direction="to">
      <uw head="fruit-tropical" />
    </icl>
  </uw>
  <relation id="uw1" label="icl" id1="uw2" id2="uw6" />
  <relation id="uw6" label="icl" id1="uw3" id2="uw7" />
  <relation id="uw3" label="agt" id1="uw4" id2="uw5" />
</sentence>

======== Deconverter Message UCL  ===========
=>Debug : [present-indicative [eat ape banane ]]

English: An ape eats a banana.
French : Un singe croque la banane.
```

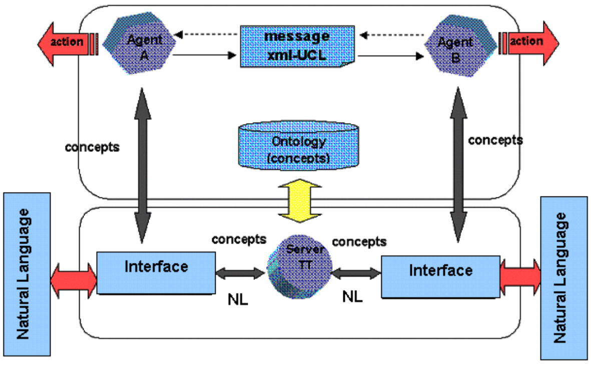


Figure 4: Architecture of a system that uses the language UCL

Figure 4 illustrates the use of UCL (using one TT server) in the communication process between two software agents.

CONCLUSIONS

The definition of the Universal Communication Language (UCL) includes all theoretical concepts of the Universal Networking Language (UNL). This was done to preserve the representative power of this language. The Web community currently regards XML as an important step toward semantic integration. Developing the language UCL using XML yielded some important benefits. The most important is the reuse of existing tools for creating, transforming, and parsing UCL documents.

The UCL enconverter-deconverter prototype shows the need for a shared ontology for the implementation of a successful enconverter-deconverter. UCL was developed to be used as a rich Agent Communication Language (ACL), which would make it easier for humans to communicate with and program software agents (using multiple natural languages). But UCL can be used in the same role as UNL.

The prototype also points out the need for an open shared ontology for UNL. UNL relation and attributes labels have some ontological knowledge already embedded in them. This makes impossible to map all possible UNL (and consequently UCL) constructs into ThoughtTreasure ontology based representation. The prototype can not be expanded into a full featured UCL enconverter-deconverter. For the time being this prototype is good enough to help the development of a prototype UCL interpreter for software agents.

The full power of the approaching of using UCL as an ACL and programming tool for software agents will only be realized, when an open shared ontology for UNL and enconverters-deconverters for many natural languages (using this shared ontology), are available. One will be able to program a software agent using his own native language and share this program with many other people, which will see and interact with the program in their own native languages.

Finally, UCL is still a proposal, but we hope that others in the Web community will help to shape its final format.

ACKNOWLEDGEMENTS

The authors would like to thanks the CNPq – National Council for Research in Brazil for the financial support for this work.

REFERENCES

Connolly, D. (2000). *Extensible Markup Language (XML)*.February 2000. Available on-line: http://www.w3.org/XML/

Dignum, Frank; Greaves, Mark. (ed.) (2000). *Issues in agent communication*. – (Lecture notes in computer science; Vol 1916: Lecture notes in artificial intelligence) Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer, 2000.

Grosof, Benjamin N.; Labrou, Yannis (1999). *An Approach to using XML and a Rule-based Content Language with an agent communication Language*. IBM Research Report. RC 21491 (96965), 28 May 1999, Available on-line: http://www.research.ibm.com

Mamadou, T. K.; Shimazu, A.; Tatsuo, N. (2000). *The State of the Art in Agent Communication Languages*. Japan Advanced Institute of Science and Technology., Japan, 1999.

Mueller, Erik T. (1998). *Natural Language processing with ThoughtTreasure*. New York: Signiform. Also available on-line: http://www.signiform.com/tt/book/

Ushida, H.; Zhu, M.; Senta, T.D. (1999). *The UNL a Gift for a Millennium*. UNU/IAS, November 1999, ISBN:4-906686-06-0.
Also available on-line: http://www.unl.ias.unu.edu/publications/index.htm

UNL, Challenges and Misunderstandings. Some answers

Edmundo Tovar*, Jesús Cardeñosa*

* Validation and Business Applications Research Group
Artificial Intelligence Department
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, Boadilla del Monte, 28660 Madrid, Spain
etovar@fi.upm.es, carde@fi.upm.es

Abstract

The UNL, either as language or as a system, is not well known due to several reasons. At this moment, UNL is not only the name of a language for computers aiming at supporting multililingual services in Internet. It is also a system, with a defined architecture and a wide panorama of applications and possibilities to support business, institutional and educational applications, all of them going beyond the linguistic barriers. Nobody doubts about the possibilities of this type of system. However, this system today supported by an organization based on a Foundation (placed in Geneva and created only to support this UN initiative) needs the collaboration and financial help of all kind of sources (UN is not financing the initiative at this moment). This is a hard task. Perhaps the more significant case about the impossibility to reach financial support for this initiative and also about the different research and application issues has been the high number of project proposals made for different Call for Proposals of the EC in the area of Econtent and IST, as well. All of them have been rejected, thus creating a wall for the development of the systems of the European languages, that are actually the more advanced within the UNL Consortium. This paper will try to analyze the different evaluations made for the various proposals in order to clear up the real state of this system and also the reasons of the low level of knowledge about this important initiative. Our goal then is to examine the opinions of the EC evaluators giving in the paper the adequate answers to them even by the side of the UNL Consortium too. Dissemination policies and internal organization of the UNL will be clarified for a better analysis in the immediate future.

1. The New Organization for the UNL Program

The Universal Networking Language (UNL in the following) arises by the initiative from the Institute of Advanced Studies (UNU/IAS), of the United Nations University in 1996. The mission of the UNL Programme (UNDL, 2002a) is to develop and promote a multilingual communication platform for Internet, with the purpose of enabling all peoples to share information and knowledge in their native language. The IAS first selected a group of institutions from fifteen countries that were in charge of developing the modules for each corresponding language. The milestones and partial results are revised in yearly meetings. In Brussels, 1999 it was presented more than a project. It was presented an Organization for the future. But the most important fact is perhaps that, after three years of development work of the participating teams, the UNL Language specifications were made public. That is, anyone can develop, potentially, for public or commercial use, components and systems integrable with the UNL system.

The current organization of the UNL Programme (UNLP) has been put on the hands of the recently created UNDL Foundation (UNDL, 2002b), a non-profit international organization created to continue the research and development initiated by the UNU/IAS in Information Technologies, in particular, in the field of interlingua communication, UNL and its applications in all areas of human knowledge and activities. The UNDL deploys the UNL facilities to assist all peoples in bridging the digital and cultural divide, in accordance with the principles and objectives of the United Nations and its Members States.

The new organization of the UNL Programme has a network of persons and institutions under the direction and sponsorship of the UNDL Foundation. The UNLP network consists of the UNL Center, the Language Centers (LC) of each language and other elements as UNL National Units, Permanent Committees and ad-hoc working groups.

The UNL Center has the overall responsibility (UNDL, 2002c) for promoting and coordinating the UNLP. In this organizational structure the LC are considered as an expansion of the UNL Center for a given language and is responsible for the research, development and maintenance of the UNL System in that language. For all these tasks, each LC must, as opposed to the past, procure the necessary financial means for the support of the LC and UNLP operations (article VII.4.e, UNLP Statute). In this paper we explain the efforts addressed by European LC, in the last years trying to get funding from the European Commission (EU) Research Programmes.

2. History of the proposals presented

European LCs have been involved in the last years in the preparation of a lot of proposals in different EU Programmes including UNL as technological base. All of them failed. We think that there are no only specific reasons to reject every proposal. This total coincidence may can be explained additionally by global reasons related to the perception and knowledge of the UNL technology by the EC evaluators as the political and technical actions of the UNL Organization.

Globally speaking, we have collected information of 8 project proposals to the EC programmes with the UNL technology as basis; all of them rejected. However, we are going to focus in this analysis over the last 5 proposals because two reasons: they are the last ones, and we, this research Group as support of the Spanish LC, has had a very active participation in them. The three other proposals were made by consortiums coordinated by companies. In all cases the application of the UNL technology fit very well with the system proposed. These proposals were named HEREIN-ML, and AQUITRA with applications for the International Office of Water. Both proposals were thought for the multilingual support of the

European Heritage Project (www.european-heritage.net). Today the European languages are covering part of the multilingual support with the UNL technology with a high degree of precarity and with direct investment in resources from the Spanish government and the Spanish Language Centre, using resources of free collaborators. However in these proposals the UNL was not accepted at all as alternative for the multilingual support.

The 5 proposals considered (see table 1), in chronological order, will be described through the summary of goals and the action lines of the Programme in which the proposal has been developed.

Date	Identifier	Programme
July 00	QANET 3312	Econtent Preparatory actions
October 00	QANET IST2000-28568	RTD Proposal, IST
January 01	LINGWEB 30130	RTD Proposal, IST
July 01	EU-UNL: 22045	Econtent Programme. Demonstration Project
November 01	MULTIDOC: 34702	RTD Proposal. IST

Table 1: A view of the proposals presented

QANET-3312(econtent): Quality Assurance Procedures for an Internet Multilingual System

Summary of goals: This proposal aimed to make a selection of resources for testing and measures of coverage together with the definition of a common lexicon of general purpose, to address the definition and construction of tools to verify and maintain their resources, to test the cross lingual tools and resources, and to generate the Quality Manuals according to ISO9001 and validation procedures to support the implantation of the Quality System in the UNL Programme.

Action lines of the RTD Programme:

Action Line 3 This proposal fits with this action because this action line mentions explicitly the problems derived from linguistic diversity and from the services to be supplied with an effective infrastructure in order to sooth the consequences of the growing number of languages in Europe and the increase of institutional and commercial relations.

QANET (RTD) IST2000-28568: Quality Assurance System for Internet Multilingual Applications

Summary of goals: This proposal aimed to define the Quality System of the UNL Program of the United Nations to overcome the linguistic barriers in Internet. The definition of the Quality System required the generation of resources (lexicon corpora addressed to this task) and methods

to evaluate the future existing systems integrated with to the UNL system. For that two industrial applications would have been developed. One UNL editor based on existing analysers (demonstrating so the integrability of this approach and the capability to reuse existing systems) and also a target language Generator completely developed from the current public specifications of the UNL language that has real intention to be exploited after the project.

Action lines of the RTD Programme:

Action Line 3.3.3 (2000 WorkProgramme). Multilingual communication services and appliances.

This proposal fits with this key action. It is addressed concretely to services and appliances independently of the language of the user. The core of the system (the UNL System) has been developed to the wider multilingual capabilities system built until this moment. In fact, UNL forms the way to access from any language and generate any other language in the world.

LINGWEB (IST2000-30130): Multilingual Web-site Deployment based on an Interlingua Technology.

Summary of goals: We aimed to deploy the UNL technology that up to that moment we had had several basic components to create the UNL network addressed to support the multilinguism in Internet. Concretely we aimed to obtain a complete implantation of multilingual services in the user web-site based on UNL technology. The user would be the Organisation Barcelona2004. Besides, we would define the materials and contents to support international training courses including the testing and adaptation of the tools involved in the maintenance of the system and the UNL coding like part of the Technology Transfer process. These tools should be adapted for any other uses out of the Language Centers environment.

Action lines of the RTD Programme:

Action Line 3.3.4 (2000WorkProgramme). Trials in multilingual e-service and ecommerce. The proposal is conceived as a Trial. Technology users are at the same time suppliers of the contents to offer an information service in a highly multilingual environment. What underlies the concern of this proposal is the effective evaluation of a new and open technology in a real environment, and to define accurately Technology Transference tasks. EU-UNL was presented as a Demonstration project to prove the economic viability of the service on a specific field.

EU-UNL: European Use of UNL

Summary of goals: EU-UNL focuses in the implementation of the UNL technology for multilingual dissemination of contents on the field of the quality of water and on the field of tourism. The project includes the user corporate web implementation of a multilingual document generation system and definition of procedures for technology transfer, planning and implementation of measures and cost evaluation, as well as the complete set of materials to assure the necessary support for internal training in the organizations. EU-UNL constitutes just a case-study for the viability of the implementation of this technology that can be extended to different languages and different fields.

Action Line of the econtent Programme: Action 2. Enhancing content production in a multilingual and multicultural environment. The overall goal in this action is to investigate and experiment with new strategies, partnerships and solutions for designing e-content products and services that can accommodate local languages and cultural conventions. EU-UNL aimed to demonstrate the capabilities of the UNL technology for multilingual dissemination of e-contents as well as for introducing a new paradigm of creation and management of multilingual web-sites.

MULTIDOC-IST-34702: A system for multilingual document dissemination

Summary of goals: The goal of this proposal was the development and integration of the components for a multilingual dissemination system in the Web using the public and open UNL. This technology constituted the base of the multilingual support for the proposed application. Initially we planned to demonstrate it in a workplace/business scenario, but were equally applicable in a personal dissemination scenario. For providing multilingual functionalities to Internet publishers UNL would be embedded into their current documents. We would also provide the tools needed for processing the new multilingual documents.

Action lines of the RTD Programme:

Action Line III.3.1 (2001WorkProgramme). The Multidoc directly addresses most of the concrete objectives listed under action line III.3.1, such as the advance towards a fuller realisation of the multilingual Internet for personal development and informational purposes, wider availability and more effective production and use of multilingual information, Multilanguage design, authoring and publishing of online (web) multimedia documents, or multilingual generation.

3. The evaluation of the proposals

We aim in this section to reflect the view of the proposers and the EC evaluators for each proposal described. Before, the evaluation criteria and the process followed in the IST Programme are explained.

3.1. Evaluation criteria for the IST Programme

A number of evaluation criteria are common to all the programmes of the fifth framework programmes. Independent experts examine each eligible proposal against these criteria. The specific programme decisions provide further details of these criteria and may also provide for additional evaluation criteria that apply only to the particular programme(s) concerned. Any particular interpretations of the criteria to be used for evaluation and any weights and thresholds to be applied to the criteria are set out in the programme-specific annexes to this document and referred to in calls and all relevant supporting documentation.

For the detailed examination of proposals against the criteria set out in the rules for participation, the experts will generally provide marks and comments. In addition, the experts are asked to examine certain evaluation criteria by answering a set of questions relevant to the specifications referred to in the call. The following questions are addressed at an appropriate moment in the evaluation:

- Does the proposal address the parts of the work programme, including policy issues, open for the particular call? If the proposal is only partially in line with the call, does it have sufficient merit to be considered in its entirety or partially?
- Have relevant ethical issues been adequately taken into account in the preparation of the proposal; is the proposed research compliant with fundamental ethical principles, if relevant? Is the research proposed in line with Community policies, if relevant; have appropriate safeguards/impact assessment regarding Community policies (e.g. environment) been taken into account, where necessary?
- Does the proposal follow the requirements for presentation (notably requirements for anonymity)?

In the case of negative answers to these questions, the experts are required to provide comments to justify their answers. On the basis of the experts' remarks, the Commission reserves the right not to continue with the evaluation of any proposal which is found not to fulfill one or more of the above requirements. In clear-cut cases (for example, a proposal which addresses a research task which is not open in the particular call), a proposal may be ruled out of scope or contrary to clearly stated policy requirements at the moment that the eligibility checks are carried out.

All eligible proposals that conform to the requirements of the call are examined for their quality and relevance by the Commission assisted by external experts. Experts examine proposals and provide marks for the criteria set out below (which are drawn from the decisions on the framework programmes and the "rules for participation" decisions and grouped into five main blocks). In addition, they also provide an overall mark for each block of criteria (unless a proposal fails any thresholds – see below). Experts are required to provide comments to accompany each of their marks in a form suitable for providing feedback to the proposers. These comments must be consistent with any marks awarded.

The blocks of criteria to be applied by all programmes are as follows (EC, 2001):

Scientific/Technological quality and innovation

- The quality of the research proposed and its contribution to addressing the key scientific and technological issues for achieving the objectives of the programme and/or key action;
- The originality, degree of innovation and progress beyond the state of the art, taking into account the level of risk associated with the project;
- The adequacy of the chosen approach, methodology and work plan for achieving the scientific and technological objectives.

Community added value and contribution to EU policies

- The European dimension of the problem. The extent to which the project would contribute to solving problems at the European level and that the expected impact of carrying out the work at European level would be greater than the sum of the impacts of national projects;
- The European added value of the consortium - the need to establish a critical mass in human and financial terms and the combination of complementary expertise and resources available Europe-wide in different organisations;
- The project's contribution to the implementation or the evolution of one or more EU policies (including "horizontal" policies, such as towards SMEs, etc.) or addressing problems connected with standardisation and regulation.

Contribution to Community social objectives

- The contribution of the project to improving the quality of life and health and safety (including working conditions);

- The contribution of the project to improving employment prospects and the use and development of skills in Europe;
- The contribution of the project to preserving and/or enhancing the environment and the minimum use/conservation of natural resources.

Economic development and S&T prospects

- The possible contribution to growth, in particular the usefulness and range of applications and quality of the exploitation plans, including the credibility of the partners to carry out the exploitation activities for the RTD results arising from the proposed project and/or the wider economic impact of the project;
- The strategic impact of the proposed project and its potential to improve competitiveness and the development of applications markets for the partners and the users of the RTD results;
- The contribution to European technological progress and in particular the dissemination strategies for the expected results, choice of target groups, etc.

Resources, Partnership and Management

- The quality of the management and project approach proposed, in particular the appropriateness, clarity, consistency, efficiency and completeness of the proposed tasks, the scheduling arrangements (with milestones) and the management structure. In addition, the tools to be used for monitoring project progress, including the quality of specified indicators of impact and performance, and ensuring good communication within the project consortium;
- The quality of the partnership and involvement of users and/or other actors in the field when appropriate; in particular, the scientific/technical competence and expertise and the roles and functions within the consortium and the complementarity of the partners;
- The appropriateness of the resources - the manpower effort for each partner and task, the quality and/or level and/or type of manpower allocated, durables, consumables, travel and any other resources to be used. In addition, the resources not reflected in the budget (e.g. facilities to carry out the research and the expertise of key personnel). For this criterion, comments may be given rather than marks.

When examining proposals, experts will only apply these criteria, supplemented by any programme-specific criteria contained in the programme decision. These criteria as they apply to the particular programme may be described in greater detail in the programme-specific annex and the work programme. Experts are not be allowed to apply criteria which deviate from those set out and the programme-specific annex.

3.2. The Evaluation of the UNL Proposals

3.2.1. Evaluation Results of QANET (econtent)

The opinion of the EC Experts.

This proposal caused a good impression because, in opinion of the evaluators, showed a well documented overview of the subject, an extensive workplan, the consortium consisted of outstanding relevant experience, with a proposal well structured. However, it presented an R&D approach rather than a feasibility demonstration of econtent, as was required by the present call. For this reason the proposal fell outside the scope of the econtent call. The evaluators recommended the submission of the proposal to a more suitable EU programme.

The opinion of the EU-UNL Consortium.

We accepted the opinion of the expert evaluators.

Actions taken by the UNL partners.

We considered the evaluation of the proposal in an optimistic way. For this reason we decided to remake the proposal to be adapted to the next call of R&D IST Programme incorporating at least a company and a user (new QANET proposal).

3.2.2. Evaluation Results of QANET (RTD)

The opinion of the EC Experts.

This proposal failed to reach the threshold score on two of the criteria.

- *Scientific/technological quality and innovation.* In opinion of the EC experts the proposal did not provide a convincing integration of both aspects, quality assurance in multilingual applications and developing resources for the UNL platform. The detailed study of the state of the art in Machine Translation and NLP systems evaluation had several omissions and did not bring forward clear conclusions.
- *Economic development and S&T prospects.* A commercial partner was willing to take up the exploitation of the project results, but these were highly conditional on the success of the UNL approach. The viability of which was questionable. Likewise, the potential for commercial exploitation of Quality Assurance methodologies for Human Language Technologies is not demonstrated, and would have required a much deeper market analysis than provided in the proposal.

The opinion of the QAnet Consortium.

We proposed to develop a series of resources (corpuses and controlled dictionaries) to be produced during the project as the base of the testing of the UNL Quality System as well as to any other NLP. For this the results are not completely conditional on the success of the UNL approach. The conclusions derived from the state of the art, maybe not enough described, are that we need to produce instantiated quality models for human language technology applications (purpose of this project).

Actions taken by the UNL partners.

We decided to carry on presenting a new proposal.

3.2.3. Evaluation Results of LINGWEB

The opinion of the EC Experts.

This proposal was judged ineligible. The reasons were because:

(1) *Non-existence of technology.* Multilingual website creation technology based on UNL does not exist while it should be a prerequisite for a trial project;

(2) *Excessive resources for development.* A high level of development and integration of new components consumes more than a half of resources;

(3) *Non-study of benefits.* The benefit of the approach chosen even in terms of productivity enhancement or the impact on the management of the lifecycle of multilingual documents is not at all addressed;

(4) *Market study insufficient.* The market perspectives are not convincing despite the intention of the coordinating partner to spin-off the results.

However, as the evaluators said in their report, the idea of using UNL as an interlingua for multilingual website creation is attractive and could be reconsidered in the framework of future generation multilingual web activities.

<u>The opinion of the EU-UNL Consortium.</u>

In this occasion, we felt very surprised by the way of the rejection of this proposal (ineligible) and the reasons that explained this decision. We answer to every one of the arguments previously described:

(1) *Non-existence of technology.* The UNL technology was officially presented in UNL annual meetings at Brussels and Geneva previously to this proposal with attendance of representatives of the EC.

(2) *Excessive resources for development.* There is not any new component in this proposal. According to the requirements of the Call for this proposal we proposed the adaptation of resources and components already existing. For this task we planned 6 man month of the total 75 mm. The rest of the tasks are assigned to produce methodologies.

(3) *Non-study of benefits.* There is a whole workpackage (wp5) that addresses specifically the definition of metrics and methods for evaluating the technical and business performances and its associated costs.

(4) *Market study insufficient.* This is more subjective argument. We proposed several exploitation strategies based on the creation of new Language Centers, the promotion for the creation of new companies from the results obtained of some Business Plan made by the coordinator of the proposal, the expansion of the use of the UNL technology without costs to institutions, segmentation of the market uses, professional training for individuals that are working in the field of translation, the creation of a commercial version of the system at low price for individuals, forming associations for the developing of specific components and/or joint exploitation of specific contents with commercial interest, and by last, through the expansion of number of languages as priority.

In summary, we did not understand and we did not agree with this qualification of proposal "ineligible". What kind of political attitude of the Programme responsible were taken?

<u>Actions taken by the UNL partners.</u>

We collected the last comment of the evaluators concerning to the idea of using UNL as interlingua for multilingual website as an attractive idea and, in spite of the strong hit we received, we kept on our efforts promoting a new proposal in the econtent Programme (EU-UNL proposal).

3.2.4. Evaluation Results of EU-UNL

<u>The opinion of the EC Experts.</u>

This proposal was considered as an interesting approach to the development of an interlingua for the automatic translation of text. However, UNL, in opinion of the experts was not sufficiently established and proven. It bears too many risks and should probably addressed under an R&D Programme. They had serious concerns about UNL, hand-encoding and its long term viability. The overall score was 2 (fair). In brief, the evaluators appreciate good technical knowledge in consortium, and they think that based on this partnership this could be a good research project.

<u>The opinion of the EU-UNL Consortium.</u>

The purpose of the project is to prove the cost-effective feasibility of the integration of a well-proven translation system to a content provider deployment strategy. This would provide a big amount of information in several languages that would serve as base of the knowledge needed. By this reason, one of the main objectives of the proposal included a Methodology for the implementation of the multilingual UNL system, including the testing phase. Effectively, the basic components of the UNL system have been already developed in the latest years. Now, they need to be tested in an integrated way and in real environments to fine-tune the interrelation of every language components such as was planned in the proposal.

<u>Actions taken by the UNL partners.</u>

We followed the recommendations of the evaluators and we promoted a new proposal in a RTD Programme (Multidoc proposal).

3.2.5. Evaluation Results of MULTIDOC

<u>The opinion of the EC Experts.</u>

This proposal only failed to reach the threshold score on one of the criteria.

Scientific/technological quality and innovation. In opinion of the EC experts the innovative value is low as this approach to the translation is not new. The scientific value of the proposal rests on the merits of the technology, UNL, that is being applied. But for these experts UNL is not a well-proven translation system since it is not backed up by solid independent evidence. Thus, this proposal fails to adduce any reference in the literature in support of UNL. By other side, according to their opinions, the proposal does not contain any suggestion how the enormous linguistic complexities of the encoding process can be taken one stage beyond machine aided / validated human effort which renders the approach economically unviable on any scale.

The overall conclusion is that the project intends to employ a technology of insufficiently proven feasibility and questionably economic viability.

<u>The opinion of the Multidoc Consortium.</u>

We propose to use the UNL technology for representing the informal contents of web pages following the XML-compliance of document mark-up languages. It is true that is not innovative. The innovative aspect in this project is the design and implementation of a multilingual dissemination system that covers all the steps of the publication chain: encoding of contents, generation of multilingual count parts and delivery of language specific versions to readers. The user site and the sites of the technology providers engage in a communication process involving standardized UNL-enriched documents using Internet-based communication software components.

As regards the complexity of the encoding process, in this moment several partners of the consortium have prototyped tools addressing this need.

Actions taken by the UNL partners.

We decided to take a period of reflection. We have taken a lot of man-month dedicated to the elaboration of proposals for the IST Programme without success. This is not a problem of a proposal but the perception of the UNL technology by the EC responsible.

3.2.6. Comparative Analysis of Evaluation Results

We have gathered the scores provided by the evaluators for previous proposals (see table 2). Each column corresponds to the scores obtained by each criterion, with the following meaning:

- Criterion 1: Scientific/technological quality and innovation
- Criterion 2: Community added value and contribution of EC policies
- Criterion 3: Contribution to Community social objectives
- Criterion 4: Economic development and S&T prospects
- Criterion 5: Resources, partnership and management

Proposal	Score Crit.1/3	Score Crit.2/2	Score Crit.3/0	Score Crit.4/3	Score Crit.5/2
QANET	Non numerical score. Global score = 0 (rejected)				
QANET	2	3	2	2	2
LINGWEB	Ineligible				
EU-UNL	Non numerical score. Global score = 2 (fair)				
MULTIDOC	2	3	3	3	2

Table 2: A view of the proposals evaluation

We have included in the table, together with the identifier of criterion, the threshold score required by the EC. An analysis of these results for the previous evaluations shows that the main obstacle for the approval of the proposals refers to the use of the UNL as technology (criterion of the technological quality and innovation). Evaluators do not find attractive and feasible the inclusion of this technology. However, in these proposals, the other criteria are in general well considered, issues such as the adequacy for the problem that address and its contribution to community social objectives, the fitness to the EC policies or a consortium balanced.

4. Conclusions

The initiatives described in this paper show at least two issues by the side of the European UNL LC (proposers). Firstly, proposers have shown a persistent interest to involve the EC in the success and diffusion of a technology for Multilinguality derived from the United Nations. Second, proposers have dedicated lot of resources trying to follow the recommendations of evaluators. Specifically, the Spanish Language Center was the coordinator of the first three proposals and was an active contributor to the rest. We have commented the last five proposals, but there are another three presented with the same results: HEREIN-ML (Towards a methodology for making textual information about European heritage multilingual by using UNL as metadata), AQUITRA and COACH (Company Organization for Automation Customer Help Integrated into ebusiness).

The diagnosis has been done but there are no clear causes. We can speculate with some of them.

- From the viewpoint of the EC evaluators, UNL technology is not feasible maybe because the lack of successful experiments and by the scarce presence of UNL in scientific areas of the sector.
- From the viewpoint of the proposers, we regret the absence or extension of more explanations or advices for the future, maybe at the political and strategic level in order to avoid apparent contradictions in the specific evaluations obtained.

The only view of all this information placed together is speaking by itself. All the proposers have long and intense European projects experience during the last ten years at least. On the other hand, it is not understandable why the interest of the EC in this global initiative of the UN is so low or inexistent. Europe must not be out of this initiative and some of the technical evaluations seem to be made in the best case by persons with a low level of knowledge about this initiative. The reader of this paper can extract conclusions by him/herself according the proposals, and the persistent and sometimes contradictory evaluations of all of them.

5. References

EC, 2001. *Manual of Proposal Evaluation Procedures*, IST Program, ed. 1-10-2001, http://www.cordis.lu/ist.

UNDL Foundation, 2002a. *The Universal Networking Language Programme. Mission*, http://www.undl.org/missionunlp.html.

UNDL Foundation, 2002b. *UNDL Foundation. Mission*, http://www.undl.org/mission.html.

UNDL Foundation, 2002c. *The Universal Networking Language Programme. Statute*, http://www.undl.org/statuteUNLP.html.