

# Overcoming the Language Barriers in the Web: The UNL-Approach

Munpyo HONG    Oliver STREITER

## 1. Introduction

In this paper we introduce the UNL Project and address the possibilities of taming existing linguistic resources into the current purpose, i.e. to develop a UNL-Enconverter for German.

### 1.1 The UNL Project

The UNL-project is a long-term project related to the storage, retrieval, exchange and presentation of information throughout the Internet. As a backbone of this project, a world-wide network of universities and research institutes from 14 countries has been set up under the guidance of the Institute of Advanced Studies (IAS) at the United Nations University (UNU) in Tokyo. In Germany, IAI at the University of Saarbrücken is the official UNL project partner with DFKI as its sub-contractor.

### 1.2. Universal Networking Language (UNL)

Universal Networking Language (UNL) developed at UNU is a formal language for representing the meaning of natural language sentences. This language is assumed to express meanings in the same standardized way as HTML presents its layout.

A UNL expression is a (possibly) cyclic graph composed of nodes connected by semantic relations. Nodes, or Universal Words (UWs) are words loaned from English and disambiguated by their positioning in a knowledge base (KB) of conceptual hierarchies. Function words, such as determiners and auxiliaries are represented in the form of attributes to UWs, provided that these function words contribute to the meaning and are not syntactically motivated. In some cases, e.g. for modal auxiliaries,

UNL allows for more than one coding strategies, i.e. to featurize a word, or keep it as a concept.

Each relation is labeled with one of the possible label descriptors. Relations that link UWs are labeled with semantic roles of the type such as *agent*, *object*, *experiencer*, *time*, *place*, *cause*, which characterize the relationships between the concepts participating in the events or states a natural language sentence may denote.

```
attrib(red(icl>state).@present.@not.@topnode,car(icl>vehicle).@def.@topic)
attrib(small(icl>state),car)
```

Figure 1: A simplified UNL expression

A simplified example of a UNL expression, as given in figure 1, shows the different components of a UNL expression. Each line represents a binary relation between UWs: *red* and *car* on one hand and *small* and *car* on the other hand. These relations are linked into a graph via the common element *car* (two identical UWs represent one concept as long as their referential difference is not explicitly marked). The semantic type of the relations is ‘attrib’, the relation between an object and its attribute. UWs may be further specified by reference to the KB (in our example *red* and *small* are described as a ‘state’ and *car* as a ‘vehicle’). Additional semantic specification can be added via the ‘@’ operator. The feature ‘topnode’ finally is used to distinguish the sentence *The small car is not red* from *The car which is not red is small*.

### 1.3 The UNL Scenario

As a result of this standardized meaning representation, documents no longer need to be multiplied in order to represent the content in different natural languages. The meaning representation is directly available to retrieval and indexing mechanisms and tools for automatic summarizing and knowledge extraction, and it will be converted to a natural language only when communicating with a human user.

The task of the presentation of a UNL web-page to a web user will be taken over by a UNL-viewer. In one commercially oriented scenario,

the UNL-viewer represents a new generation of web-browsers which, in addition to their capacities to handle Java and Java-script, are equipped with one or more national UNL-Deconverter in order to display the meaning content in a natural language. In the scenario of distributed NLP<sup>1</sup>, the web-browser is linked to a national site of the UNL Network and this national site performs the task or subtasks of decoding. In this latter scenario, the national servers could be equally equipped with a huge bulk of encoded examples, so that the decoder can be supported with approaches similar to Translation Memories and Example Based MT systems.

The UNL-documents to be made available in the Internet are prepared neither manually nor fully automatically. The formal and linguistic specifications of this language are far too complex to be fulfilled by an untrained and unsupported person. Therefore, the creation of UNL-documents is supported by national UNL Enconverters which convert a natural language text in a raw version of UNL. This raw version is to be visualized and edited in UNL-editors, a tool currently developed by the UNL-network to be used by trained UNL writers to finalize the UNL document.

## 2. The CAT2 MT-System in UNL

CAT2 is a unification-based NLP formalism developed for the purpose of multilingual MT. Within this rule-based formalism different grammars and lexicons have been developed by different project groups. The system developed currently at IAI is linked to an Example-Based MT component<sup>2</sup> and incorporates statistical knowledge related to different subject domains<sup>3</sup>, so that the whole preprocessing is currently supported by three integrated MT paradigms.

---

<sup>1</sup> Cf. SCHUBERT (1988) and STREITER et al. (1998).

<sup>2</sup> Cf. CARL et al. (1999).

<sup>3</sup> Cf. STREITER et al. (1999).

## 2.1 The German-to-UNL Enconverter

The general strategy followed in the development of a German-to-UNL Enconverter is firstly to use the existing modules of the MT system as much as possible. Secondly, the system should supply the user with as much information as possible, especially in the case of a failing analysis. Thus, even if no complete UNL expression can be created for a German sentence, the system should present at least the equivalent UWs, their attributes, and those binary relations between the UWs that could be identified.

After morphological analysis, the German syntactic analysis is performed by the kernel of the UNL Enconverter, the CAT2 German-to-English MT-System introduced above. In the analysis phase we avoid the time-consuming complex analysis in a deep level but try to build a simple syntactic structure by a limited number of construction-specific rules; there are specific rules for certain types of linguistic constructions such as passive, modal verb construction, present- and past-perfect construction, etc. All these modules remained unchanged for this UNL-related task.

## 2.2 Generation of UWs

The most important task for the UNL Enconverter is the identification of the correct UWs and their positioning in the conceptual hierarchy. This subtask has to be performed in any case, even during robust processing. Although UWs are supposed to be language independent concepts, the strings of the UWs are identical to English words. By no means, however, a UW is the English word with the same string, nor does the usage of an English word motivate the usage of this word as UW. English colloquial words or slang words, for example, are represented in UNL by their standard English equivalent. At the moment, however, UNL does not possess a feature system to systematically represent the speech style<sup>4</sup>, the tenor (the speaker-hearer relation) nor the 'channel of communication'<sup>5</sup>,

---

<sup>4</sup> Cf. DE MAURO (1994).

<sup>5</sup> Cf. BELL (1991).

so that such properties of natural language expressions simply disappear. In order to generate UWs, the German-to-English MT-System can be used. It possesses a very flexible translation mechanism, where not words but Notional Domains (NDs) are linked in transfer.<sup>6</sup> According to this mechanism, every term of the target language, e.g. *kontrollierbar* (controllable), is included into a German ND of KONTROLLIEREN (control). This ND is linked in transfer, among others, to the English ND of CONTROL. A comparison of the semantic and stylistic properties of *kontrollieren* with those terms available in the ND CONTROL yields a set of possible translations (e.g. *control*, *controllable*, *controlling*, *control*). The syntactic context of these terms helps to select a unique English word. This selection is further refined by features of markedness, including a mechanism preferring the shortest variant of a target word (e.g. preferring an irregular verb-to-noun derivation to a regular verb-to-noun derivation). Within this transfer mechanism, it has been easy to add a feature into the description of an English word which describes, if all the preceding mechanisms do not help to exclude a word, this word as an illegal UW. In the following simplified example we show how the German colloquial expression *betucht* is matched onto the UW *rich*:

```
{lex=reich}&
({slex=reich,head={ADJ}}
;{slex=betucht,ct={style=colloq},head={ADJ}})&
{trans={en=({t=(rich;abound;abundant;affluent;wealth)})},
```

Figure 2: Simplified lexical entry matching a German colloquial expression onto standard English words, i.e. potential UWs

In order to handle the wide range of subjects which can be found in the Internet, the integration of a statistical component into the transfer component has become necessary. This statistical component helps to select the ND which is most likely to be used in a given subject domain (e.g. ‘medicine’, ‘sports’, ‘computing’ etc). For this purpose, the source

---

<sup>6</sup> Cf. STREITER & SCHMIDT-WIGGER (1995) and STREITER (1998).

language lexicon (German) is supplied with the frequency information concerning English words in different subject domains. If the German ND containing the German source word is linked to more than one English ND as shown in figure 2, the system will try the ND which has the highest frequency within this subject domain first. By this means, the German translation of the word *Tor* is first looked for in the ND of *goal* in the sport context, and if this does not fit, in the ND of *gate*. Since the statistical data are based on monolingual corpora, some negative effects of this approach cannot be totally avoided. These negative effects and a more effective approach are discussed in STREITER et al. (1999).

### 2.3 Generation of Graphs and Labels

In the step of generation of English equivalents, the necessary UNL-features are inserted. These features are (a) lexical features which specify the UW and (b) features which describe the relations between UWs. The lexical features of the UWs come (i) out of the English Lexicon (e.g. the positioning in the conceptual hierarchy) (ii) out of the English syntacto-semantic context (e.g. information with respect to the question whether a word is the axiom of the structure) and (iii) from the transfer (e.g. information about tense, aspect, modality, determination, quantification, mood etc.). While (ii) depends on the success of the English generation, (i) and (iii) depend only on the degree to which the German source sentence could be parsed and transformed to the usual CAT2 Interface Structure. This is exemplified below for the German sentence *Das kleine Auto ist*

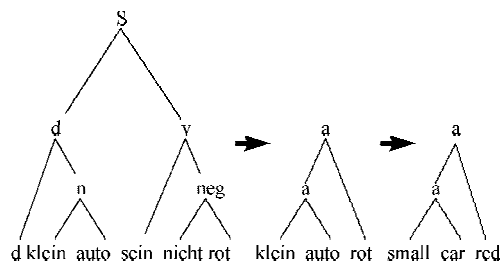


Figure 3: Syntactic analysis of a German sentence and transformation into the English Interface Structure

*nicht rot* (the small car is not red). Once the analysis is completed, lexical transfer into English takes place, so that the German words are replaced by their English equivalents. The result of this transformation is a normal English CAT2 Interface Structure.

For information coming out of the English lexicon, simple lexical rules are sufficient. These rules translate the existing lexical information into the appropriate UNL-descriptors. The semantic information of words, for example, is stored in CAT2 in the SEM-feature and the process type (PROC). This feature structure is translated via a special mapping rule into a corresponding semantic type in the knowledge base of UNL.

```
f_activity = {proc=act,sem={evt={stat=no}}} >> {unl={sem=activity}}.[].  
f_state    = {sem={evt={stat=yes}}} >> {unl={sem=state}}.[].  
f_vehicle  = {sem={ccr={t=vehicle}}} >> {unl={sem=vehicle}}.[].
```

Figure 4: Mapping of lexical CAT2 features onto UNL features

Other grammatical and semantic information coming from the source language such as the tense of a verb or the number and the definiteness of nouns is transformed into a semantic feature and transferred into the target language equivalent of the source language item. From this semantic feature, UNL attributes are calculated in a similar simple way as seen before.

```
f_def      = {reference=familiar} >> {unl={attrib1=def}}.[].  
f_indef    = {reference=unfam} >> {unl={attrib1=indefsg}}.[].
```

Figure 5: Mapping of Contextual Features to UNL

At the lexical level all English words bear the UNL features which have been calculated by these rules:

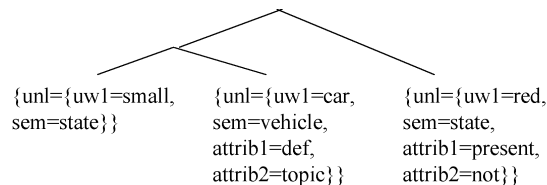


Figure 6: Lexical UNL features

The features describing the relation between UWs are inserted into the syntactic rules of English. As these rules are multiplied for different lexical entity types (*'attributive adjective – noun construction'*, *'attributive active participle – noun construction'*, *'attributive passive participle – noun construction'*), the necessary UNL-specifications can be easily added to these rules. The binary adjective – noun construction for example adds the following UNL-features into the description of the adjective:

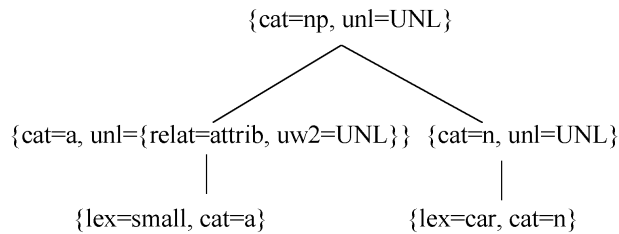


Figure 7: Creation of UNL-features in a Nominal Phrase

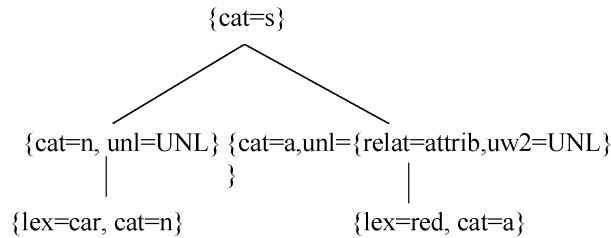


Figure 8: Creation of UNL-features in a Copular Structure

As can be seen in figure 7 and 8, every binary relation between English words is represented by a UNL-feature of one of these words. This feature, which collects all UNL-related pieces of information, describes, among others, *uw1* (the source of the graph), *uw2* (the destination of the graph), *relat* (the UNL label for this graph) etc. This UNL-feature represents the output of the MT component in place of the morpho-syntactic specifications for the morphological generator.

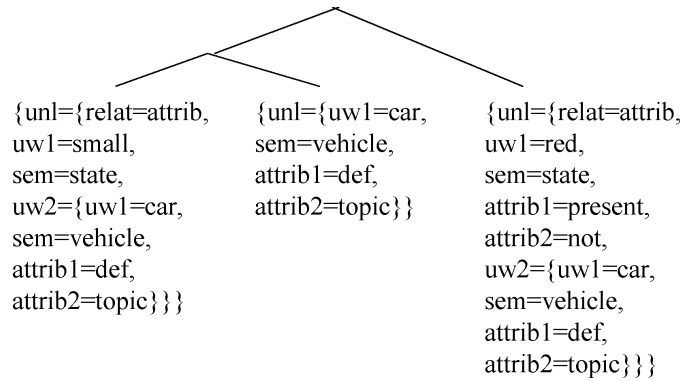


Figure 9: UNL features accumulated by the MT system

The latter is replaced by a small Perl-program which suppresses terminals which are not sources of graphs (e.g. the noun *car* in our example) and transforms the UNL-feature into one annotated UNL graph, as shown below:

```

attrib(red(icl>state).@present.@not.@topnode,car(icl>vehicle).@def.@topic)
attrib(small(icl>state),car)
  
```

Figure 10: Simplified UNL for the German sentence *Das Auto ist nicht rot*

### 3. Conclusions

In this paper we have described the employment of a Rule-based MT system within the UNL network. We have tried to show that in spite of the conceptual differences between UNL and the theories underlying MT systems, such systems can take over the tasks of UNL enconverting. UNL Enconverters share with conventional MT systems the analysis and transfer module, so that we believe most of the existing MT-systems to be easily tuned into a UNL Enconverter.

In addition, the lexical resources these systems possess and process can be employed as a Lexicon Server, in order to supply other UNL-related modules via the Internet with alternative or additional lexical codings. It is by this strategy that the German UNL Decoding is realized

at DFKI, by connecting their own German generation tool via the Internet to the German Lexicon Server located at IAI.

Test versions of the components described here can be accessed via the Internet. The German Lexicon Server which transforms UW into German lexical items annotated with all semantic and syntactic information can be found at <http://www.iai.uni-sb.de>. The German UNL Encoder can be accessed at the same site. However, in order to guarantee a smooth continuation of the project, only the UNL members retain the right to access to these tools. The above URL may equally serve as a starting point in order to know more about UNL.

## References

- BELL, Roger T. (1991): *Translation and Translating. Theory and Practice*. Applied Linguistic and Language Study. London and New York: Longman.
- CARL et al. (1999): Michael C., Catherine PEASE and Oliver STREITER, Examples of hybrid Machine Translation. In: *ISMT and CLIP*, Beijing.
- DE MAURO, Michael Tullio (1994): Sette forme di adeguatezza della traduzione. In: *Capire le parole*. Roma-Bari: Sagittari Lateranza.
- SCHUBERT, Klaus Michael (1988): The architecture of DLT – interlingua or double direct? In: Dan MAXWELL, Klaus SCHUBERT and Toon WITKAM (eds.), *New Directions in Machine Translation*. Dordrecht / Holland: Foris Publication.
- STREITER, Oliver (1998): A semantic description language for multilingual NLP. In: *The Tuscan Word Centre – Institut für Deutsche Sprache Workshop on Multilingual Lexical Semantics*, 19-21 June 1998.
- STREITER, Oliver and SCHMIDT-WIGGER, Antje (1995): Patterns of Derivation. In: *TMI95: Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Leuven, Belgium, 5-7 July 1995.
- STREITER et al. (1998): Oliver S., Antje SCHMIDT-WIGGER, Ursula REUTHER and Catherine PEASE, Experiments in distributed MT. In: *Workshop on Distributing and Accessing Linguistic Resources*, Granada, Spain.
- STREITER et al. (1999): Oliver S., Leonid L. IOMDIN, Munpyo HONG, Ute HAUCK, Learning, Forgetting and Remembering: Statistical Support for Rule-based MT. In: *TMI99: Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, University College Chester, England, 23-25 August 1999.