Standardization of the generation process in a multilingual environment

Carolina Gallardo
Dpt. of Information Organization and Structure
E.U. de Informática
Universidad Politécnica de Madrid
cgallardo@eui.upm.es

Jesús Cardeñosa
Dpt. of Artificial Intelligence
Facultad de Informática
Universidad Politécnica de Madrid
carde@ fi.upm.es

ABSTRACT
Natural language generation has received less attention within the field of natural language processing than natural language understanding. One possible reason is the non-standardization of the input for generation systems. This is an obstacle to the systematic planning of the process of developing generation systems. We propose the use of UNL as a possible standard for standardizing generation inputs.

Keywords
Natural language generation, standardization, multilinguality.

1. INTRODUCTION
Natural language processing (NLP) is divided into two areas: analysis and generation. However, the scientific community has not lent the same measure of attention to the two fields, and generation can be considered as NLP's "poor brother".

The reason for this relative underdevelopment is that analysis and generation systems have different inputs. The input for analysis systems is always natural language, which has a well-known casuistry and phenomenology. On the other hand, although we know what the output of a generation system will be, we do not, in principle, know what this output will be generated from, [1].

A generation system input varies depending on whether it generates monolingual (dialogue systems) or multilingual (mainly machine translation systems) output. In dialogue systems it is hard to set formal common input features, as the generation "problem" is usually dealt with using ad hoc solutions that depend on the application and system languages. Likewise, there is a also wide variety of inputs to the generation component of machine translation systems. They are conditioned by the type of system architecture (transfer, interlingua, etc.), the type of grammars used (declarative vs. procedural) [2], or the number of system languages.

Due to the disparity of generator inputs, it is impossible to systematically plan the generator development process (the main reason for the underdevelopment of generation compared with analysis). To do this, a formal format- and language-independent content representation model assuring a standard generation system development process is required to support the generator input.

In this paper we propose UNL as a possible standard for generation inputs. To do this, Section 2 will introduce the main generation architectures. Section 3 will describe the UNL system, its properties and baseline architecture in more detail. Section 4 will establish the conditions that a technology must meet to be able to be considered a standard and what conditions UNL meets.

2. GENERATION ARCHITECTURES
2.1 Dialogue Systems
Dialogue systems are one of the main natural language generation applications. The purpose of dialogue systems is to "present information to users in a format that they find easy to understand" [3] in very specific domains where the user and the system usually interact in the same language. The user asks the system for some information. Having gathered the required information, the system can generate a natural language response to visualize the information. In many cases, this response is built (quite successfully) by generating a constructed language from a series of templates that have a predefined relation to the templates supporting the queries [4]. In other words, the input for the generation process is very much dependent on the form of the user's original query. It could be said that there is no thorough analysis of the text, nor is there an abstract representation of the information to be presented to the user. The overdependence on the source and domain language is an obstacle to the construction of multilingual dialogue systems and the reuse of such systems in other domains.

2.2 Machine Translation Systems
Machine translation (MT) systems are inherently multilingual because their goal is to transform a text written in language A into an equivalent text in language B. In this section we will describe the main MT system architectures, as each architecture sets a series of conditions to be met by the formal properties of the generation inputs.

2.2.1. Transfer Systems
The basic tasks of a transfer system are analysis, transfer and generation. The analysis component outputs a (more or less deep) source-language-dependent syntactic representation of the input text. This syntactic representation is the input for the transfer module whose job is to transform this representation into a more target-language-like structure. The transfer module output is the system generation module input. It is this module that finally outputs the target language sentence. The components, inputs and outputs of transfer systems are strongly reliant on the source and target languages.

Even within the transfer architecture, the transfer system generation processes vary widely. This variation is accounted for by:

- *Transfer component functions*: this component is responsible for transforming the abstract representation of

the source language into a more target language-like representation. The transfer tasks could be embedded in the analysis (the analysis is very target language reliant), and the generation tasks (e.g. word reordering) are carried out during the transfer process. The immediate consequence is that there is a morphological synthesis process rather than a generation process in the strict sense. The Metal system [5] operates according to this architecture.*Analysis and generation process symmetry*: the system can opt to output a non-target-language-dependent output. In this case, the transfer module would be responsible for adapting the output to the target language. In these systems, the generation process is clearly delimited and completely independent of the transfer process. The ETAP-3 system [6] illustrates system symmetry with respect to the analysis and generation modules.

The main problem with transfer systems that do not clearly separate the transfer and generation components is that the resources (grammars and dictionaries) and components are not reusable for creating new language pairs in the system. Really, if we wanted to increase the number of language pairs in the system, we would have to build a new system. Generally, the dependency of transfer systems on a target language leads to more precise outputs but makes it harder to reuse the components to build new languages into the system.

2.2.2 Interlingua-based systems

Interlingua-based systems are the second major MT system paradigm. Interlingua-based systems include both traditional systems, like ATLAS-II [7], PIVOT [8], and knowledge-based systems, like KANT [9] or Mikrokosmos [10]. Their distinguishing features are:

- *Removal of the transfer process.* The system carries out two basic tasks: analysis and generation.
- *Single intermediate representation.* Since the transfer module has been removed, the abstract representation output by the analysis directly "feeds" the generation module. This intermediate representation is the component called "interlingua".

Interlingua-based systems are designed to cover more possible languages, as an interlingua-based system requires 2*n components for n languages. This is noticeably less than the n*(n-1) components that transfer systems need for the same number of languages.

An interlingua must satisfy a number of requirements:

- It must be a natural-language-independent representation.
- It must be capable of representing the semantics of natural languages (to enable the generation process).

Whereas the required level of abstraction in transfer systems was syntactic analysis, this level is not good enough for an interlingua-based system insofar as it would be of no use for generating other languages. A deeper representation level further removed from the *form* of the input language is required.

Figure 1 shows the basic generation architecture in an interlingua-based system.

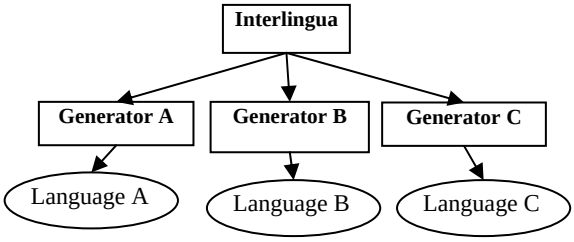

Fig. 1. Generation in interlingua systems

Interlingua-based systems have a key advantage over transfer systems: the architecture enables the inclusion of new languages and all the components are reusable. On the other hand, though, key grammatical information for generation could be lost during the interlingua conversion process, that is, the interlingua may contain less (grammatical, not conceptual) information than a syntactic representation. In short, interlingua-based systems offer more languages in exchange for less precise generated texts.

2.2.3 Fusion

Multilinguality is unquestionably an added value for any generation system. The transfer-interlingua dichotomy would suggest an opposition between precision and number of languages. In an attempt to take advantage of the strengths of the two paradigms, some transfer systems have *interlingualized* their architectures to support a greater number of languages [11],[12]. The key characteristic of these systems is that there is a detailed syntactic representation that is to some extent independent of the source language. The process for merging the interlingua architecture into a transfer system requires the construction of a transfer module between the deep-syntactic structure and an interlingua representation [13].

3. THE UNL APPROACH

3.1 The UNL system

The UNL language (Uchida, 2002) is an artificial language designed to represent the content of texts written in any natural language. UNL is furnished with some specifications that formally define the language. They are:

- *Universal Words*. UNL does not suggest a set of primitive concepts as Schank [15] or Jackendoff [16] do. In order to express a lexicalized concept in natural language, UNL proposes the use of English words modified by a series of semantic constraints that remove the ambiguity inherent in the natural language vocabularies. If there is no satisfactory English entry to express the concept, UNL permits the use of words from other languages provided that the semantic constraints precisely describe the meaning of the base word. This gives the language the expressive wealth of natural languages with none of their ambiguity. An example of a universal word would be:

  construcción$_1$ → construction(icl>action)
  construcción$_2$ → construction(icl>concrete thing)

  (where "icl" is the abbreviation for "included").

- *Relations*. They are a set of 41 relations defining the semantic links between concepts. They can be argumentative (agent, object, goal), circumstantial

(purpose, time, place), logical relations ("and" and "or"), etc.

- *Attributes*. Attributes express semantic information derived from morphological flexion and the functional elements of the sentence (auxiliary verbs, articles, etc). They are added to the universal words to further specify their meaning when they appear in a particular context. Attributes include information on the event tense or aspect, number, polarity, mood, etc.

Formally, a UNL expression takes the form of a semantic net, where the nodes (universal words) are linked by labelled arcs to UNL conceptual relations. For example, the UNL representation of the sentence "the boy eats potatoes in the kitchen" is:

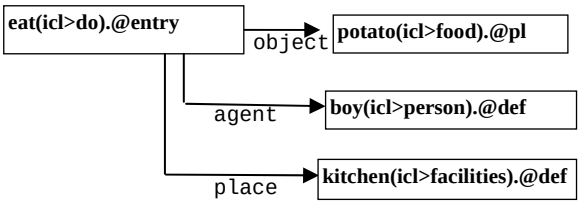

Fig. 2. Representation of a UNL expression

The syntax of the written representation of this sentence is:

agt(eat(icl>do).@entry, boy(icl>person).@def)
obj(eat(icl>do).@entry, potato(icl>food).@pl)
plc(eat(icl>do).@entry, kitchen(icl>facilities).@def)

3.2 Basic architecture of the UNL system

The UNL system represents a generic framework for the massive generation of multilingual contents. Its key goal is to represent the contents of a document, web page, database, etc., as an *approved and standardized structure* that can be transformed into natural language text. The distinguishing features of the UNL system are:

a) UNL is designed to generate multilingual contents. A document written in UNL is an "entity" in itself and can be stored in a document base, etc.
b) UNL does not imply the use of special-purpose components or tools. The tools, components and process that are defined to edit and generate UNL all vary from one language to another. The use of UNL merely implies the standardization of the generation system input [17].

Although the system places more emphasis on language generation, the UNL framework includes both the process of editing natural language in UNL (called "enconverting") and the process of generating natural languages (called "deconverting").

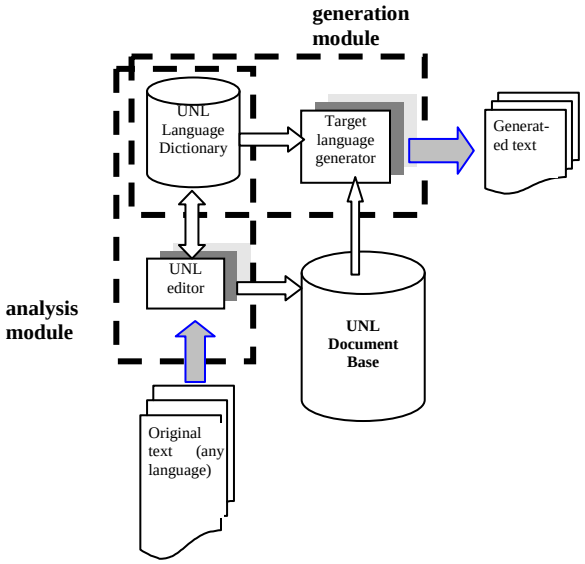

Fig. 3. Architecture of the UNL system

UNL is essentially an interlingua, that is, a formal language for independently representing the meaning of natural languages. The UNL language is not confined to a specific domain (like the KANT or Mikrokosmos interlinguas). By placing no initial constraints on the set of interlingua vocabulary, we guarantee that UNL is able to represent the contents in any language or domain.

3.3 Generation in the UNL framework

There are several architectures for generating natural language from UNL. Let us detail the two generation architectures within the UNL framework.

3.3.1. Direct generation

The UNDL centre (http://www.undl.org) provides a module for enacting the generation process as a single process. All the grammatical knowledge required to generate the target language is to be found in the dictionary and in the language-specific rule base. As a semantic representation is converted directly into a morphological implementation, the dictionary has to contain as detailed as possible information on aspects like:

- *Grammatical category and subcategories*: the quality of the generation can be expected to improve the greater the level of lexical hierarchization is.
- *Argumentative structure and prepositions governed by* verbs, nouns and adjectives.
- *Semantic information* likely to be relevant for the syntactic configuration of the target language.

With the help of the information contained in the dictionary, the main job of the generation rules is to transform the UNL expression into a natural language sentence. The tasks carried out are basically:

- Match UNL relations to language-specific grammar relations.
- *Translate* the UNL attributes to their respective morphological or lexical implementation. For example, UNL only offers three options in the case of tense. It would be the *job* of the generation rules to select the right verb tense and mood for the languages that do not conform to this tense system (e.g. Spanish).
- Generation of pronouns and anaphoric expressions.
- Morphological synthesis.

Figure 4 shows the direct generation architecture.

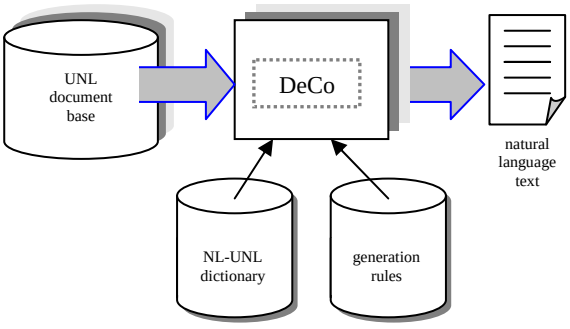

Figure 4: UNL direct generation architecture

3.3.1. Mixed generation: reuse of transfer components

The processing of the Russian and French languages within the UNL system is an example of *mixed generation* within the UNL framework. Both teams have integrated the UNL system into their transfer systems: ETAP system in the case of Russian [12] and Ariane for French [18].

These systems opted to reuse the available target language generators and to develop an additional module to convert the UNL representation into a format readable by their transfer system generators. Figure 5 shows what a mixed architecture would look like.

This, of course, involves generating a new component, generically referred to as the UNL transfer module. Even so, the experience with the above systems has shown that it costs much less to develop this module than it would to develop a new generator that accepted UNL code as a direct input.

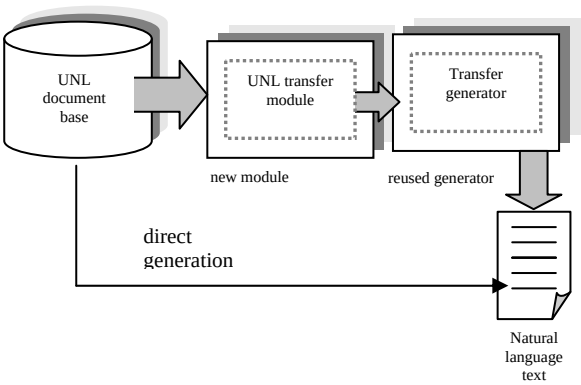

Fig. 5. Mixed generation

4. UNL AS A STANDARD FOR MULTILINGUAL GENERATION

A standard is composed of a set of obtainable criteria that are used to determine the conformity of an object or an action. This may not be a very helpful definition. But, in the case of software technologies, standards also provide guidance on how software products should be defined and also offer the possibility of benchmarking the products implemented using the standard.

UNL is a standard for supporting Internet-based multilingual services. The key to the success of any standard is the maturity concept. Maturity can refer to different aspects proper both to the technology and to the supporting organization. UNL, specifically, exploits the advantage in terms of cost cutting that interlingua systems have over the translation processes in transfer systems, especially when there are more language pairs. UNL enables both public and private institutions to provide services, such as disseminating research results, reporting regulations among divisions of a multinational, or setting up the basis for expanding electronic commerce. Also it offers further competence-related advantages: it is operational from an Internet-based distributed environment, and it is supported by the United Nations. Thanks to this, it can pursue a social purpose such as protecting minority languages, irrespective of commercial issues.

UNL has complementary resources and tools. Apart from the accessibility of the technology, which is public, it is possible to use the standard to develop different multilingual applications.

All this is supported by the United Nations' UNDL Foundation, set up specifically to exploit the technology.
These and other issues related to the standard were compared against software technologies standards that have proved to be successful. The weaknesses of UNL in this respect are due primarily to the fact that its implementation is still in its early stages [19].

5. CONCLUSIONS

It looks as if an interlingua system architecture (based on a single input format for all generators) is a good support for the idea of formally defining the input for developing the generator component in compliance with formal specifications (as yet not available). This formal specification would be the basis of a generator development standard, and would set up an environment for testing the quality of this key component in the generation of multilingual contents.

In view of UNL's language properties (natural language independence and adequacy for expressing any natural language content) and the possibility of integrating the UNL system with any existing generation system, UNL is a good candidate for a standard for normalizing natural language generation systems.

A standard should be supported by an organization that assures its stability and maintenance. In this case, this organization already exists: the UNDL Foundation's UNL Centre under the United Nations umbrella.

5. REFERENCES

[1] R. Dale, B. Di Eugenio, and D. Scott. "Introduction to the special issue on Natural Language Generation". *Computational Linguistics*, vol 24(3). 1998.

[2] P. Whitelock. *Linguistics and computational techniques in machine translation system design*. London. UCL Press in association with the Centre for Computational Linguistics. 1995.

[3] E. Reiter, and R. Dale. *Building applied natural language systems*. Cambridge University Press. 1995

[4] A. Ballim, and V. Payota. “Weighted Semantic Parsing: A robust approach to interpretation of Natural Language Queries”. *Flexible Query Answering Systems:* H. Larsen et al. (eds): Physica Verlag. 2001.

[5] W.P. Lehmann, W.S. Bennett, J. Slocum, H. Smith, S.M.V. Pfluger, and S.A. Eveland. *The METAL system*. Final technical Report, RADC-TR-80-374, Linguistics Research Center, University of Texas at Austin. NTIS AO-97896. 1981

[6] J. Apresjan, I. Boguslavskij, L. Iomdin, A. Lazurskij, V. San-nikov, and L. Tsinman. “ETAP-2: The linguistics of a Machine Translation system”. *META*: XXXVII(4):97-112. 1992

[7] H. Uchida. ATLAS-II: A machine translation system using conceptual structure as an interlingua. In Proceedings of the Second Machine Translation Summit, Tokyo. 1989.

[8] K. Muraki. “PIVOT: Two-phase machine translation system”. *Proceedings of the Second Machine Translation Summit*, Tokyo. 1989.

[9] E. Nyberg, and T. Mitamura. “The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains”. *Proceedings de COLING-92: 5th International Conference on Computational Linguistics.* 1992

[10] S. Beale, S. Nirenburg, and K. Mahesh. “Semantic Analysis in the Mikrokosmos Machine Translation Project”. *Proceedings of the 2nd Symposium on Natural Language Processing*. Bangkok, Thailand. 1995

[11] T. Aikawa, M. Melero, L. Schwartz, and A. Wu. “Multilingual Natural Language Generation”. *Proceedings of MT Summit VIII*. Santiago de Compostela. 2001

[12] I. Boguslavsky, N. Frid , L. Iomdin, L. Kreidlin, I. Sagalova, and V. Sizov. “Creating a Universal Networking Language Module within an Advanced NLP System”. *Proceedings de COLING 2000: 18th International Conference on Computational Linguistics.* Saarbrucken. 2000

[13] B. Lavoie, R. Kittredge, R. Korelsky, O. Rambow. “A Framework for MT and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing”. *Proceedings of 6th Applied Natural Language Processing Conference*. ACL, Seattle. 2000.

[14] H. Uchida. The Universal Networking Language. Specifications. http://www.undl.org. 2002

[15] R. C. Schank. *The fourteen primitive actions and their inferences*. Memo AIM-183, Stanford Artificial Intelligence Laboratory. 1973.

[16] R. Jackendoff.. *Semantic Structures.* Current Studies in Linguistics series. Cambridge, Massachusetts: The MIT Press. 1990.

[17] C. Boitet, and G. Sérasset. “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”. *Proceedings COLING 2000: 18th International Conference on Computational Linguistics*. Saarbrucken. 2000

[18] C. Boitet, and G. Sérasset. “UNL-French deconversion as transfer & generation from an interlingua with possible quality enhancement through offline human interaction”. *Machine Translation Summit 99*. Singapore. 1999.

[19] E. Tovar, and J. Cardeñosa. “A Descriptive Structure to Assess the Maturity of a Standard: Application to the UNL System”. *Proceedings of the 2nd IEEE Conference on Standardization and Innovation in Information Technology*. Boulder, Colorado. 2001.