Multilingual documents management by using
Universal Networking Language UNL on Alignment Gestion Tool OGA

Mutsuko TOMOKIYO & Abdel-Basset AL ASSIMI & Christian BOITET
GETA, CLIPS, IMAG-campus, BP53, 385 rue de la Bibliothèque, F-38041 Grenoble
tel : +33 (0)4 76 51 44 83, fax : +33 (0)4 76 51 44 05
{Mutsuko.Tomokiyo, Abboud.Assimi, Christian.Boitet}@imag.fr

Keywords

Universal Networking Language (UNL), Universal Word (UW), converter/deconverter, multilingual translation, linguistic alignment and correspondences of documents, parallel multilingual documents

Summary

The paper aims to present multilingual document alignment and translation on the platform OGA (**O**util de **G**estion d'**A**lignement, Alignment Gestion Tool) by using the UNL (**U**niversal **N**etworking **L**anguage). Essential management allowed by OGA is to create a parallel multilingual document by importing in the associated data base the different monolingual files, edited on different software, corresponding to its resources in several languages.
UNL is a sort of interlingua such as pivot languages in certain MT systems, which enables to translate documents into several different languages. UNL system contains UW dictionaries, decorder and encorder.
The typical sitution OGA engages in is the internationalization of same documents in different languages in small office, that's the situation where documents alignment and translation are required at the same time. OGA was developed to fulfill these needs in a user-friendly environment.
We describe, first, OGA and UNL language. Second, we introduce UNL expressions rewritten from French leaflet by screen images on OGA. Third, we make a French generation experimentation by using a deconverter. Finally, we discuss the adventage of UNL language and its application to the translation on OGA, while comparing OGA with other software.

Introduction

This paper aims to present multilingual document alignment and translation on the platform OGA (**O**util de **G**estion d'**A**lignement, Alignment Gestion Tool) [2] by introducing a language UNL (**U**niversal **N**etworking **L**anguage).

The typical situation OGA assumes is the internationalization of leaflets or small documents including semi-automatic translation of documents in small offices such as tourist centers, unversity departments, etc.

The general difficulties rise out of the heterogeneity of software used for same documents written in multiple languages, the absence of good softwear showing document structures, the lack of simple language representing source text in translation processing without using big and complex systems.

OGA was developed to fulfill these needs in documents processing.

In this paper , we focus on the representation of source text to be translated on the platform OGA.

First, we will give the simple explanation about OGA platform.

Second, we detail UNL language by citing some examples and show UNL expressions by screen images on OGA.

Third, we make an experimentation of French generation from UNL expressions.

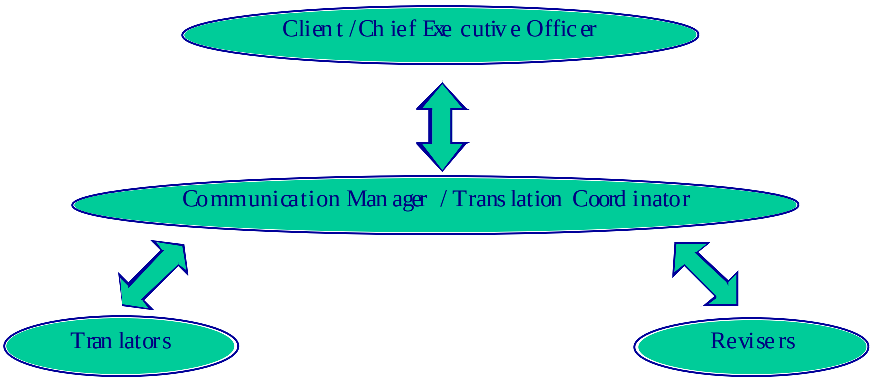
Finally, we discuss the advantage of UNL language and its application to the translation on OGA, while comparing OGA with others software.

We will use the leaflet of *IMAG Institute* as source text. The leaflet describes the organization of *Imag Institute*, its localization, research contents, directors and supervisors of each of laboratories, etc. in 8 languages : Arabic, Chinese, English, French, Japanese, Portuguese, Russian, Spanish.

1. OGA

1.1 OGA managements

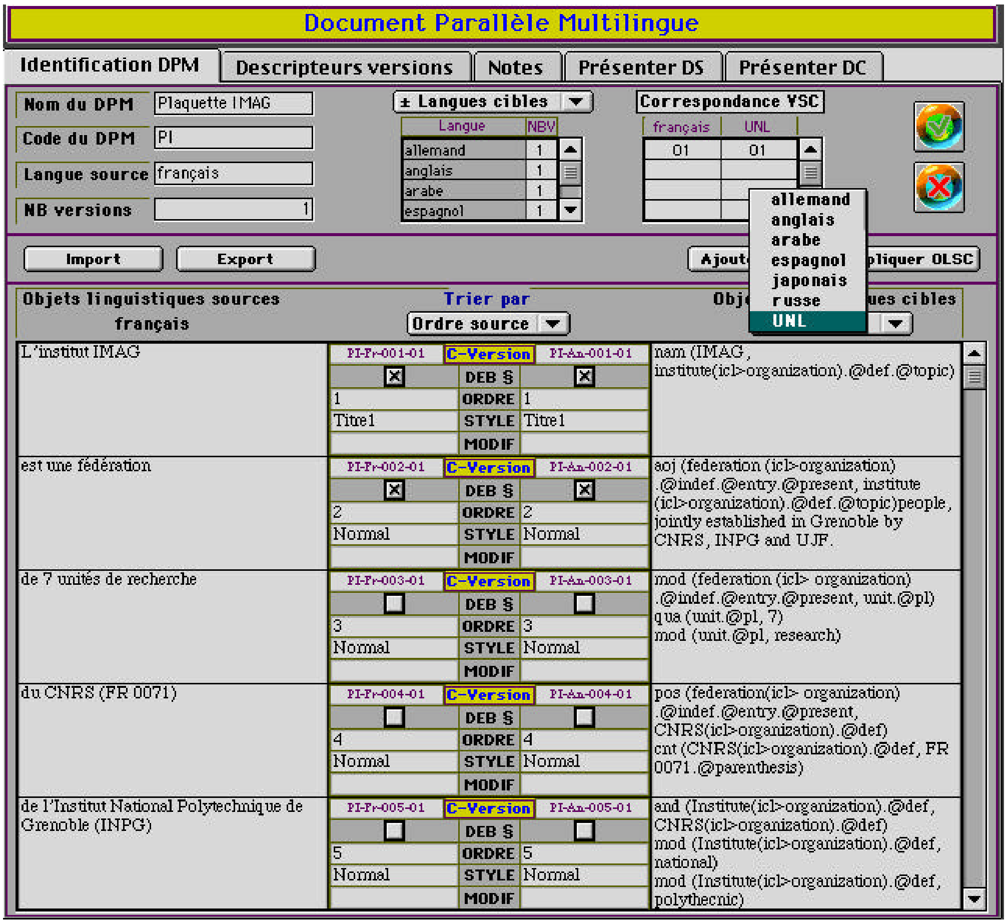
OGA is a workbench for multilingual documents alignment, and is set in order to be used interactively with users as shown in fig.1.


Client /Chief Executive Officer
Communication Manager / Translation Coordinator
Tranlators
Revisers

(fig.1)

The normal task flow is the :

1. creation of the parallel multilingual documents by importing in the associated data base the different monolingual files, originally formatted in different way with different software packages ;
2. preparation of the new original document after negociation with the client (the person in charge of the information content) ;
3. back translation of modifiered parts into source language and creation the new source document validated by the client, and revision of the new complete translated monoligual files ;
4. production of final monolingual files in XML [1] [2].



Nom du DPM: Plaquette IMAG
Code du DPM: PI
Langue source: français
NB versions: 1

Langue	NBV
allemand	1
anglais	1
arabe	1
espagnol	1

français	UNL
01	01

Trier par: Ordre source

Objets linguistiques sources français				
L'institut IMAG	PI-Fr-001-01 ☒ 1 Titre1	C-Version DEB § ORDRE STYLE MODIF	PI-An-001-01 ☒ 1 Titre1	nam (IMAG , institute(icl>organization).@def.@topic)
est une fédération	PI-Fr-002-01 ☒ 2 Normal	C-Version DEB § ORDRE STYLE MODIF	PI-An-002-01 ☒ 2 Normal	aoj (federation (icl>organization) .@indef.@entry.@present, institute (icl>organization).@def.@topic)people, jointly established in Grenoble by CNRS, INPG and UJF.
de 7 unités de recherche	PI-Fr-003-01 ☐ 3 Normal	C-Version DEB § ORDRE STYLE MODIF	PI-An-003-01 ☐ 3 Normal	mod (federation (icl> organization) .@indef.@entry.@present, unit.@pl) qua (unit.@pl, 7) mod (unit.@pl, research)
du CNRS (FR 0071)	PI-Fr-004-01 ☐ 4 Normal	C-Version DEB § ORDRE STYLE MODIF	PI-An-004-01 ☐ 4 Normal	pos (federation(icl> organization) .@indef.@entry.@present, CNRS(icl>organization).@def) cnt (CNRS(icl>organization).@def, FR 0071.@parenthesis)
de l'Institut National Polytechnique de Grenoble (INPG)	PI-Fr-005-01 ☐ 5 Normal	C-Version DEB § ORDRE STYLE MODIF	PI-An-005-01 ☐ 5 Normal	and (Institute(icl>organization).@def, CNRS(icl>organization).@def) mod (Institute(icl>organization).@def, national) mod (Institute(icl>organization).@def, polythecnic)

(fig.2)

So, OGA is not a simple aligner : aligners are used to prepare the imputting format. Rather, OGA's main technical goal is to centralize the correspondances between many versions of the same document, possibly more than 1 in each language. From the information system of view, OGA is simple enough to be used by translators or secretaries, and powerful to effect automatically reratively complex tasks.

As a springboard, we have experimented OGA for the translation and production of the leaflet of *IMAG Institute*. The following screen images show French (in the left side column) and English (in right side column) texts are aligned with some annotations such as sentence number, sentence style, modifications if happened, in the central column.

2. Introduction of « UNL » to OGA

2.1. UNL

UNL[1] is an electronic language which enables to rewrite articles in any languages on Internet into UNL format in order to translate them into any other languages, and allows to establish a communication among people of different mother tongues.

Our idea consists in introducing this language into OGA and in coping with multilingual translation for small leaflets or documents.

[1] UNL was developed by UNL center, which consists of the members of 16 countries under the aegis of the Institute of Advanced Studies (IAS) of the Organization of United Nations.
The enconverter and deconverter are supplied the member of UNL society by UNL center [7].

The UNL includes UW's dictionary (**U**niversal **W**ords dictionary) and two software : *enconverter* and *deconverter*.
E*nconverter* is an analyser applicable to any source languages, and contains analysis rules and dictionary. It enables to analyse source languages at morpho-syntactic and semantic level.
Deconverter is a generator also applicable to any target languages. Its role is morpho-syntactic and semantic generation, while using generation rules based on ontological knowledge and information on words concatenation in a sentence.
Universal Word (UW) is UNL's vocabularies. A basic UW is made up of English-like words, compound words, and phrases. To restrict the meanings of the basic UW, we attach ontological restriction labels to the basic UW.
The process of restriction or making Uws expressing narrower concepts is as followed :

(1) Choice of concept category : nominal concept (icl>thing), verbal concept(icl>do, icl>occur or icl>be), adjectival concept(aoj>thing or mod<thing), adverbial concept(icl>how)
(2) Addition of UW hierarchy of UNL KB or case relations, when the ambiguity of an UW cannot be solved : *e.g.* swallow(icl>bird), swallow(icl>action), swallow(icl>quantity), spring(icl>do(obj>wood)), spring(icl>do(obj>mine)), spring(icl>do(obj>person, src>prison))
(3) Subcategorization of UW concept : the 4 concept categories are subcategorized into *e.g.* abstract thing, concrete thing, functional thing, place, volutional thing, etc.

Thus, Uws are eliminated the ambiguity of transfer : for example, the word "*spring*" is defined as "*spring*" of a tool and "spring" of a season by indicating them by the symbol « icl> ».

e.g.
spring (icl>tool)
spring (icl>season)
institute(icl>organization)
institite(icl<do)
federation (icl>organization)
federation(icl>states)

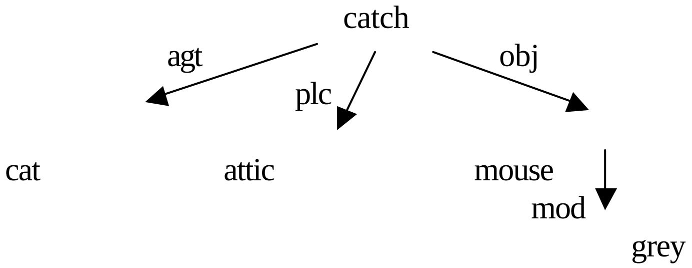
The source text is converted into a format having "relation", "attributes" and "ontological labels", and represented implicitly in hypergraph of nodes and arcs.
Ontological labels allow to define the possible relations among UWs, and eliminate possible ambiguity of the source language.

"Relation" represents the syntax in UNL's paradigm, in other words the relations between UWs of a sentence. The « relation » is described as a set of relation labels which consists of 50 relations such as agt, aoj, obj, cag, cob, gol, man, plc, src, tim, etc., based on deep case grammar [5]. They are attributed on the arcs in hypergraph.
"Attributes" are used to describe what is said from thr speaker's point of view : how the speaker views what is said. This includes phenomena tecnically called "speech acts", "propositional attitudes", "truth values", etc. [8, page 18/23]. They are associated to the nodes of a graph.

e.g.
The cat catches a grey mouse in the attic.
catch – agt-> cat.
catch – obj -> mouse
catch – place-> attic

This means that "cat" is the actant of the action "catching", and the object of the action is "mouse". When this sentence is converted into UNL format, the following representation is obtained :

agt (catch.@entry.@present, cat (icl>animal)@.def)
obj (catch.@entry.@present, mouse (icl>animal).@indef)
plc (catch.@entry.@present, attic.@def)
mod (mouse (icl>animal).@indef, grey (icl>color)   (UNL_expression 1), [3]



In the UNL_expression 1, « @entry » indicates entry point or main UW of whole expression or in hyper node, and « @present » defines speakers narrative time as present. « @def/@indef» means that the noun phrase is marked as definitive or indefinitive.

Thus, when UWs dictionary is prepared, all languages will be able to be converted into UNL format.

2.2. IMAG Institute leaflet into UNL format

We converted *IMAG Institute* leaflets in French and Japanese into UNL format. The UNL expressions followed are the same part of the leaflet as we saw in the left side column of fig.2

Source text in French	annotations	UNL expressions
L'institut IMAG	Sentense number 1 [p] [s]	nam (IMAG, institute(icl>organization).@def.@topic)
est une fédération		aoj (federation(icl>organization).@indef.@entry.@present, institute(icl>organization).@def.@topic)
de 7 unités de recherche		mod (federation (icl> organization).@indef.@entry.@present, unit.@pl) qua (unit.@pl, 7) mod (unit.@pl, research)
du CNRS (FR 0071)		pos (federation(icl> organization).@indef.@entry.@present, CNRS(icl>organization).@def) cnt (CNRS(icl>organization).@def, FR 0071.@parenthesis)
de l'Institut National Polythechnique de Grenoble (INPG)		and (Institute(icl>organization).@def, CNRS(icl>organization).@def) mod (Institute(icl>organization).@def, national) mod (Institute(icl>organization).@def, polythecnic) nam (Institute(icl>organization).@def, grenoble) cnt (Institute(icl>organization).@def, INPG.@parenthesis)
et de l'Université Joseph Fourrier (UJF)	[/p] [/s]	and (University(icl>organization).@def, Institute(icl>organization).@def)

		nam (University(icl>organization).@def, joseph(icl>first)) nam (University(icl>organization).@def, fourrier(icl>family)) cnt (University(icl>organization).@def, UJF.@parenthesis)

(UNL_expression 2)

[p] = beginning of paragraph, [s] = beginning of sentence, [/s] = end of sentence, [/p] = end of paragraph

The UNL expressions for the leaflet of *IMAG institute* in any languages are quite same excepte for its UWs. This is thanks to UNL language which converts the concept of source text into target language. When a source text in a language has converted into UNL expressions, the conversion of multilinguages is easy, as long as UW dictionaries is developed for each languguage in order to generate source texts.
The test of French generation for the *IMAG* leaflet by the deconverter[2] is shown Appendix [4].

3. Discussion & Conclusion

Our experimentation in documents alignment and translation shows strong feasibility of personal translation on OGA. The UNL system consists of simple configuration, UW dictionary, converter and deconverter. UNL language specification is based on the deep case grammar, which is close to human intuition. OGA itself shows rigour capacity of source text format recognition.
On the one hand, there is some alignment software, for example « Wordfaster » of Champollion, which uses strategies of style recognition and terminology recognition as pure alignment software. The software's purpose is essentially to generate translation memories from pairs of translated documents. They are a set of segmented units of a document being a pair of one source segment matched to its corresponding translation, and are created automatically by marking language name and sequential number of the segmented units [9]. « Wordfaster » cannot deal with files formatted in different ways by different software.
The other hand, there is « Org-explorer » developped by UNL center as worldwide news database written in UNL exressions. This offers a user-friendly Windows platform that permets to search by type organization and type of information, but this is not an alignment system [6].
In the managements on OGA, translation is combined with alignment processing, and it's done interactively with final users. So validations in one by one manner of works are promised.

[2] This deconverter was developed independently from UNL center in the laboratories CLIPS-GETA-IMAG.

Bibliographie

[1] *Al Assimi A.-B. & Boitet Ch.*, Management of non-centralized evolution of parallel multilingual documents, the 10th International WWW Conference, Hong Kong, 2001

[2] *Al Assimi A.-B*, Gestion de l'évolution non centralisée de documents parallèles multilingues.

[3] *Blanc,E.*, From the UNL hypergraph to GETA's multilevel tree, Proceedings of MT2000, BCS, 2000

[4] Sérasset G., Improuving Lexical Transfer in UNL French Decnverter. Rapoort technique, GETA-CLIPS-IMAG, 1999

[5] *Tomokiyo M.,Nédeau N. & Boitet Ch.*, Merging a strictly deep-case approach and a multilevel approach in a pivot language definition for MT. Proc. Pacific Association for Computational Linguistics, Tokyo, Japan, 1997

[6] *UNL center*, Organization explorer (Org-explorer), leaflet, UNU/IAS, Genève, 2000

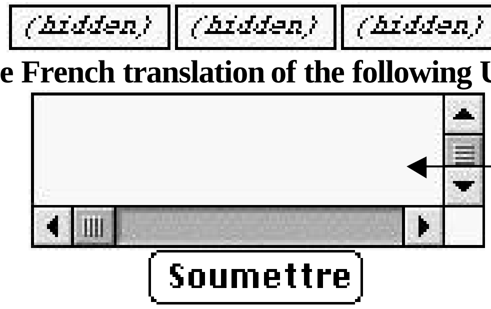
[7] http://www.unl.ias.unu.edu/info/

[8] *http://www.ias.unu.edu/research_prog/science_technology/universalnetwork_language.html*

[9] *Yves Champollion and the Wordfast development team, Paris 2,* WordAlign Manual, http://champollion.net, 2000

Appendix

DECO gate - UNL to French Deconverter



Obtain the French translation of the following UNL graph:

Input : [p]

[s]

nam (IMAG, institute(icl>organization).@def.@topic)

aoj (federation(icl>organization).@indef.@entry.@present,

:

cnt(Institute(icl>organization).@def, INPG.@parenthesis)

[/s]

[/p]

Output : L'institut <<IMAG>> est une fédération de 7 unités d'une recherche ( ) l'institut d'un ressortissant <<POLYTHECNIC>> <<GRENOBLE>> (<<INPG>>) l'université <<JOSEPH>> <<FOURRIER>> (<<UJF>>) .