A "Pivot" XML-Based Architecture for Multilingual, Multiversion Documents: Parallel Monolingual Documents Aligned Through a Central Correspondence Descriptor and Possible Use of UNL

Najeh Hajlaoui, Christian Boitet

équipe GETA, laboratoire CLIPS,
385 rue de la bibliothèque - BP 53, 38041 Grenoble Cedex 9 - France
Christian.Boitet@imag.fr

Abstract. We propose a structure for multilingual, multiversion documents, built on the model of the web-oriented, cooperative lexical multilingual data base PAPILLON: a document is represented by a collection of monolingual XML "volumes" interlinked by a central volume of "interlingual links". Here, the links relate subdocuments (XML trees) corresponding to each other in monolingual "volumes". We are developing a Java application to enable direct editing of a multilingual document through the web, at the level of monolingual volumes as well as through bilingual or trilingual interfaces inspired by those of commercial "translation workbenches". Another goal is easy integration with machine translation and multilingual generation tools. For this, we add a special UNL volume. In a first stage, we split the UNL-xml document in several monolingual documents, again represented by XML files. Each document contains the text in a particular language, plus the corresponding UNL graphs, and can be modified independently. The interface is easy to build, but realigning the documents after a series of such modifications is a very difficult task.

1 Introduction

Due to Internet, the number of available documents grows dramatically. There is a strategic need for companies to control information written in more than 30 languages (HP, IBM, MS, Caterpillar). This requires the installation of powerful and effective management tools of multilingual "synchronized" documents.
There are techniques of large-grained linking (on the level of HTML pages). However, there are no techniques for structuring multilingual documents so as to allow fine-grained synchronization (at paragraph or sentence level) and even less permitting editability through the Web.

The interest to synchronize at least on the level of the sentences is double:

– for the translation and human revision with the assistance of techniques of HTHM (Human Translation Helped by Machine) and in particular of translation memory.

Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 309–325.

– for the increase in the number of languages of a multilingual document, it would be useful to synchronize the versions of multilingual documents with a representation such as the multilingual UNL document format, allowing to increase the number of languages of the document in an economic way by calling distant deconverters.

The paper is organized as follows.

In the first part, we put our research in perspective with the UNL project (Universal Networking Language). We show the advantage and the limits of the UNL format, and discuss some aspects related to the management of the information systems.

In the second part, we present a possible solution to manage the correspondences between the linguistic versions of a multilingual document: it consists in splitting a document in UNL format in several monolingual documents.

The third part is devoted to the reconstitution of links broken between the documents, and to a mockup and prototypes of interfaces.

In the conclusion, we show the flexibility of such a structure of multilingual, multiversion documents, and its applicability in several domains.

2 Problems

2.1 Situation of the Problem

There are many multilingual documents, which are modified separately (leaflets, booklet, etc.). After a certain time, we wish to make them coherent [1]. That means finding the correspondences (alignments) and reconstituting a complete and coherent (monolingual) "source" document. For this, modifications in target languages have to be translated into the source language.

A. Assimi, in his PhD work, treated the case of the non-centralized management of the evolution of multilingual parallel documents.

In the industry, it is frequent that documents are managed on the same platform without being linked at a fine-grained level like that of sentences or paragraphs.

For example, technical documents are usually aligned at the level of HTML pages. Generally, free modification by readers (final users) is not authorized (whereas it is usually permitted for leaflets in Word).

Several problems appear in real life:

1. As shown in Figure 1, alignment (based on sentences considered to be exact mutual translations of each other) may be quite sparse, even with only 2 languages, after only one batch of modifications in one language.
2. There is no explicit link between the monolingual (real) documents constituting the (virtual) multilingual document.
3. In some contexts like the European Heritage web site, a UNL document is also built in parallel, as a simple list of UNL-graphs, with no document structure. The problem can then be abstracted as in Table 1 below.

L'institut IMAG	IMAG institute
est une fédération de 7 unités de recherche du CNRS, de l'Institut National Polytechnique de Grenoble (INPG) et de l'Université Joseph Fourier (UJF). L'IMAG représente une communauté de 650 personnes (dont la moitié de doctorants) qui se consacre à la formation et à la recherche en informatique et mathématiques appliquées.	
	The Computer Science and Applied Mathematics Institute of Grenoble (IMAG) accounts for most of the academic research in these domains in Grenoble. IMAG is a federation of seven laboratories, comprising about 650 people, jointly established in Grenoble by CNRS, INPG and UJF.
Depuis 1988, l'Institut IMAG est l'interlocuteur des tutelles, des collectivités territoriales et des industriels ou institutions avec lesquels il mène des partenariats pluriannuels ; coordonne et anime la vie scientifique inter et supra-laboratoires : mise en évidence de projets de recherche soulignant les axes scientifiques de l'Institut, projets d'expérimentation avancée, formations doctorales, colloques et écoles ; gère les ressources communes aux différents laboratoires : réseau et moyens informatiques, médiathèque, services électronique et infographie, cellules communication et multimédia, affaires internationales.	
	These laboratories have a long standing tradition of cooperation with industry and of active participation in European programs. They may be credited with an indisputable ability to apply their results and transfer their know-how from research to industry.
Un enseignement supérieur de pointe	Top level university training
Les scientifiques de l'Institut IMAG participent à la formation de plus de 1 000 étudiants de second et troisième cycle de l'ENSIMAG (école de l'INPG) et de l'UFR d'Informatique et Mathématiques Appliquées (UJF).	
	IMAG university training higher education is given each year to 1500 students by members of IMAG (professors and researchers) in one Engineering School of INPG (ENSIMAG), one University Department of UJF (UFRIMA), and in six other joint graduate schools..

Figure 1: Example of alignment.

Language 1 (FR)	Language 2 (EN)	……	UNL
φ^{FR}_{1}	φ^{EN}_{1}		γ_1
φ^{FR}_{2}	ϕ		γ_2
ϕ	φ^{EN}_{3}		γ_3
…			
φ^{FR}_{n}	φ^{EN}_{n}		γ_m
…			
$\varphi^{FR}_{N\text{-}FR}$	$\varphi^{EN}_{N\text{-}EN}$		γ_M

Table 1: correspondences between sentences.

φ^{l}_{i} = sentence with identifier i in language l

γ_m = *UNL* graph representing the meaning of one (occurrence of a) sentence

A very simple idea is to seek an identifier for a set of equivalent sentences, with

- $\varphi_1 \cong \varphi_2$ *if and only if UNL* (φ_1) = *UNL* (φ_2)
- $\varphi^{l}_{i} \cong \varphi^{l'}_{i'}$ *if and only if* $\sigma(\varphi^{l}_{i}) = \sigma(\varphi^{l'}_{i'})$

  σ is the equivalence of the intuitive means but testable by a human translation
- $\varphi^{l}_{i} \cong \varphi^{l'}_{i'}$ *if and only if* $\rho(\varphi^{l}_{i}) = \rho(\varphi^{l'}_{i'})$

  ρ is defined in a restrictive and operational way. Here, ρ = UNL.

A first problem is to calculate links from the UNL graphs to the sentences in each monolingual document. They may be modeled as a function $\Pi : 1..M \times L \rightarrow \mathbb{N}$ or as a relation in $1..M \times L \times \mathbb{N}$.

If we choose the first possibility, a UNL graph (in the parallel UNL document) cannot be linked by Π to more than 1 sentence in any language, which implies that 2 identical UNL graphs can appear in the central list. The idea is that, after some reordering and duplication, the list of UNL graphs can be linked to the list of sentences (the terminal nodes) of the xml structure of each monolingual document, with "no crossing". In other words, Π is then monotonically increasing in its first component.

We might also choose the second possibility, where Π is a relation, so that all occurrences of sentences with the same meaning could be linked to the same UNL graph. Then, the parallel UNL file should represent a set of UNL graphs, with no possible repetition.

However, both these possibilities lead to problems. Let us show it on the first only (Π is a function). Then,

$\Pi (m, l) = n$ if and only if

1. $\delta (\gamma_m, l) \approx \varphi^l_n$ where δ stands for "deconversion" (from UNL)
2. $\lambda (\varphi^l_n) = \gamma_m$ where λ stands for "enconversion" (into UNL)

$\Pi (m, l) = \underline{\text{nil}}$ otherwise (γ_m does not correspond to any sentence).

To establish the links between the UNL graphs and the sentences implies then to call all deconverters on all graphs, and to compare the results with the actual sentences. But deconverters are constantly updated, may be unavailable at some time, and sentences may also be modified by hand. Hence, with all probability, only very few links will be established. What would be needed is a process to compare the meaning of a sentence present in a document with that of a sentence produced by deconversion "on-the-fly". But that is a hard and perhaps harder problem!

We can also attack the problem from the other side, that is, we can try to establish links from the sentences to the UNL graphs. This linking is the inverse ψ of Π. Again, ψ can be a function or a relation. In the UNL format, it is a function, which implies that, if a sentence is truly ambiguous and corresponds to several different UNL graphs, one of them has to be chosen in the representation. Let us adopt this restriction.

We have then $\psi : \mathbb{N} \times L \rightarrow \mathbb{N}$, and

$\psi (n, l) = m$ if and only if $\Pi (m, l) = n$.

We encounter a similar problem: to compute ψ, we have to "enconvert" each sentence, and compare the result with the UNL graphs in the list. But (1) enconversion is harder than deconversion, and (2) the UNL language allows for more than one way of representing a given interpretation of a sentence.

We should then develop techniques to test the synonymy of 2 UNL graphs… but it is quite certain that any proposed solution will be incomplete, because the problem of deciding whether 2 formal expressions have the same meaning is undecidable as soon as the considered formulas pertain to a rich enough formal system. For example, it is undecidable whether 2 java programs compute the same function.

This shows that the solution consisting in putting some UNL-related or UNL-like representation as a central structure leads to problems. It also imposes the added difficulty to build a correct and complete UNL-xml document.

Hence, our solution will be to design a specific central structure linked to all sentences of all monolingual documents, and to the UNL graphs. A separate problem will be to determine whether some intersection or union of the monolingual document structures should be reflected in the central structure or not.

2.1.1 Evolution of the Versions of a Multilingual Document

We introduce the term "polyphrase" to denote a set of sentences in several languages and UNL graphs, formed from an initial set of such elements, the "kernel" of the polyphrase, deemed to be semantically equivalent.

In most cases, the kernel is simply one sentence in a given language, all other sentences are obtained by translation or corrections, and the UNL graphs by enconversion and then direct edition or coedition.

While the kernel corresponds to exactly one intended meaning, the evolution of the polyphrase may introduce new meanings. To trace them, we need to add a notion of version to the elements of a polyphrase, and by extension to all parts of a multilingual document.

The passage to a new version can happen in many cases:

- correction of errors.
- human revision.
- addition of another language.
- change of order of linguistic objects.
- addition of new polyphrases.

The preceding points are important factors, which influence the increase in the number of versions of a multilingual document, and the unalignment rate of these versions.

2.1.2 Coherence of the Versions

The coherence of the versions is directly related to the concept of alignment. 2 versions in 2 languages will said to be "coherent" if their aligned documents are mutual translations of each other. Alignments should go at least to the level of sentences. In our first mockup (see below), we stop there, but finer units such as segments and words may be quite useful to help human translators or posteditors.

The coherence of the versions of the database is distinct from that of an environment of translation; the graph of dependence is fixed and the ascending translation process respecting alignment generates a coherent version.

A new version then traverses a development cycle until it becomes frozen and/or validated, before entering in a state of "public" availability. It can then be used in a translation memory.

2.2 Advantages of the UNL Language and Limits of the UNL Format

We choose UNL [2] as our interlingua for various reasons:

1. it is specifically designed for linguistic and semantic machine processing,
2. it derives with many improvements from H. Uchida's pivot used in ATLAS-II (Fujitsu), still judged as the best quality MT system for English-Japanese, with a large coverage (586,000 lexical entries in each language in 2001),
3. participants of the UNL project[1] have built "deconverters" from UNL into about 12 languages, and at least the Arabic, Indonesian, Italian, French, Russian, Spanish, and Thai deconverters were accessible for experimentation through a web interface in spring 2003,
4. although formal, UNL graphs (see below) are quite easy to understand with little training and may be presented in a "localized" way to naive users by translating UNL symbols (semantic relations, attributes) and lexemes (UWs) into symbols and lexemes of their language,
5. the UNL project has defined a format embedded in html for files containing a complete multilingual document aligned at the level of utterances, and produced a "visualizer" transforming a UNL file into as many html files as languages, and sending them to any web browser.

The UNL representation of a text is a list of "semantic graphs", each expressing the meaning of a 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.

The lexical units, called Universal Words (UW[2]), represent (sets of) 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 special symbol (number…) possibly completed by semantic restrictions: the UW "process" represents all word meanings of that lemma, seen as citation form (verb or noun here), and "process(icl>do, agt>person)" covers only the meanings of processing, working on, etc.

The attributes are the (semantic) number, genre, time, aspect, modality, etc., and 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 in language L is to say that it represents the abstract structure of an equivalent English utterance "seen from L", that is, where semantic attributes not necessarily expressed in L may be absent (e.g., aspect coming from French, determination or number from Japanese, etc.).

The UNL format, whether UNL-html or UNL-xml, gives for the moment a simple solution: a multilingual document is only one large file where the alignment of the various versions (languages and revisions) is done at the level of each sentence. But, in general, two parallel documents in two different languages cannot be aligned at this level. Indeed, a sentence in L1 can correspond to two or three sentences in L2 and conversely (m-n possibility).

[1] http://unl.ias.unu.edu

[2] Universal Word, or Unit of Virtual Vocabulary

Moreover, the order of a list of sentences or paragraphs can vary from one language to another (for example because of a lexicographical sorting). Thus, the idea from where we left in the introduction is good, but must be refined.

2.3 Aspects Related to the Management of Information Systems

The problem of management of correspondences and coherence of MPDs (Multilingual Parallel Documents) still remains open: there is no adequate concrete solution, indeed there is a lack of tools, methods, practices and models to describe, maintain and refine the correspondences between versions of the same document in several languages.

An important point is that the suggested techniques must be usable in practice and as practical as possible in the known information systems. Let us see how the problem is posed on this level.

2.3.1 Centralized Management

In the case of centralized management of documents, the problem is easier to solve as soon as (1) a unique XML format is used for exchanging and storing data, and (2) there is a central place to describe and control the correspondences between linguistic versions. The disadvantage, however, is that the freedom to modify individual versions is limited.

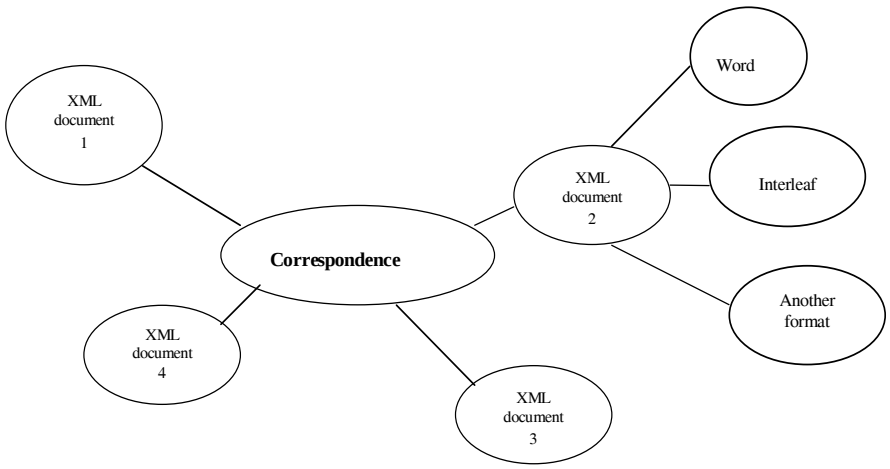


Figure 2: correspondence between centralized documents (XML formats)

Indeed, the life cycle of a multilingual document organized in this way has to be controlled from the start using certain mechanisms of observation and protection of the correspondences.

2.3.2 Non-Centralized Management

There are many cases where the various versions of a document are not centralized, for instance because they have to be processed with different tools on different platforms. To realign them after a series of modifications have been done is quite difficult, and to rebuild a coherent complete original is even more difficult.
On the formatting side, there are m-n possibility of correspondences for several distributed documents, n different formats and 2n filters.

A. Assimi [1] analyzed and solved a part of these and other problems posed by the management of the non-centralized evolution of multilingual parallel documents. He used a structuring of the multilingual texts by a multicolumn table, which is not practicable for documents of big size (technical documentation, catalogues...). In his thesis, he reports that this simple solution worked for certain needs of customers, but was limited to the management of small documents such as the brochure of the IMAG Institute (Informatics and Applied Mathematics at Grenoble), which contains approximately 2000 words, that is 8 standard pages of translation, or 4 pages of Word.

2.3.3 Principe of Solution

Starting from the study made in the two preceding cases, we see the need for designing tools and methods allowing practical management of large multilingual documents. In particular, it is necessary to describe and to maintain linguistic correspondences at a very fine level between n versions in m languages, while allowing new versions to appear in any language independently of others.

For that, the idea is to represent the correspondences between the structural trees of n parallel monolingual documents by a separate structure, of a different type, connecting fragments of trees with as few constraints as possible, as is done for the macrostructure of the multilingual lexical data base PAPILLON.

3 The Versioning Problem and a First Solution

We simply adopt the solution implemented in PAPILLON (storage of the modifications on standby in the form of XSLT transformations in the private space of each contributor) and draw from our preliminary experiment in management of versions for XML documents representing virtual electronic components.
In order to manage the successive versions of a multilingual document, we introduce the concept of status of a version.

3.1 Status of Documents and Versions

The status of any part of a document can be:

- **modifiable**: when its contents can still undergo modifications.
- **frozen**: when its contents cannot be modified but are not yet validated.

– **validated**: when its contents have been validated. A validated part may be put on some sharable reference space.

We define the order: modifiable < frozen < validated.

Suppose a multilingual document has content in n languages (including UNL if present).

The “last version” of any part of this document is the n-uple consisting of the maximum version number of all its polyphrases.

A “version” of a document is any n-uple of version numbers less or equal to the last version (component by component).

The status of a version of a document is the minimum of the statuses of the sub-document corresponding to that version.

3.2 From a Multilingual Document to Several Monolingual Documents

The basic idea is to separate the monolingual documents and to represent their correspondences in an autonomous "pivot" structure. It was also the idea of A. Assimi, but we use it here in a context where the formats to be synchronized are standard XML formats. We find it too in PAPILLON, where each dictionary of lexies (word meanings or monolingual acceptions) is represented by an XML file, as well as the "pivot" or “hub” formed by the axies (links between lexies).

In addition, more and more annotations are introduced into documents for various applications (IR, summary, categorization...). They can be annotations related to the language (like GDA of K. Hashida) or annotations only related to the contents (graphs UNL, semantic categories...).

At this point, we consider two ways of separating the monolingual documents: partial separation and total separation.

3.2.1 Partial Separation

Let us suppose for the moment that we have a multilingual document in UNL-xml format aligned on the level of the sentence. Suppose we want to switch to the non-centralized management situation, for example to let 15 persons edit the same document in 15 languages.

The idea of partial separation is then to split the UNL-xml representation into 15 monolingual documents enriched by the original content (source language) and its UNL representation as shown in the following example.

This makes it possible to make local modifications in each language and thus to introduce different versions. Here, for example, the sentence "He eats fruits" becomes "He is eating fruits" with the corresponding modification of UNL-xml format, and a second version of the English document appears (figure 4).

3.2.3 Total Separation

Here, we split the UNL-xml representation in several monolingual documents by considering the fact that the original is also a monolingual document as well as its UNL representation.

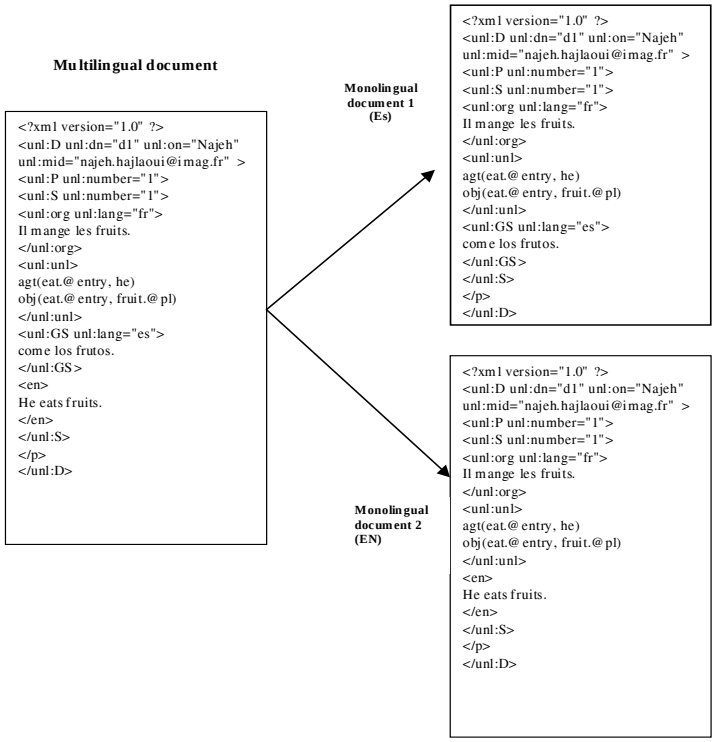


Figure 3: partial separation of a multilingual document.

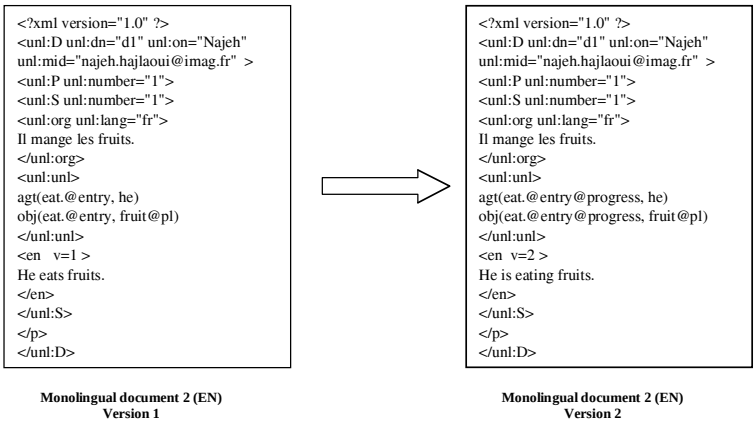


Figure 4: evolution of monolingual document.

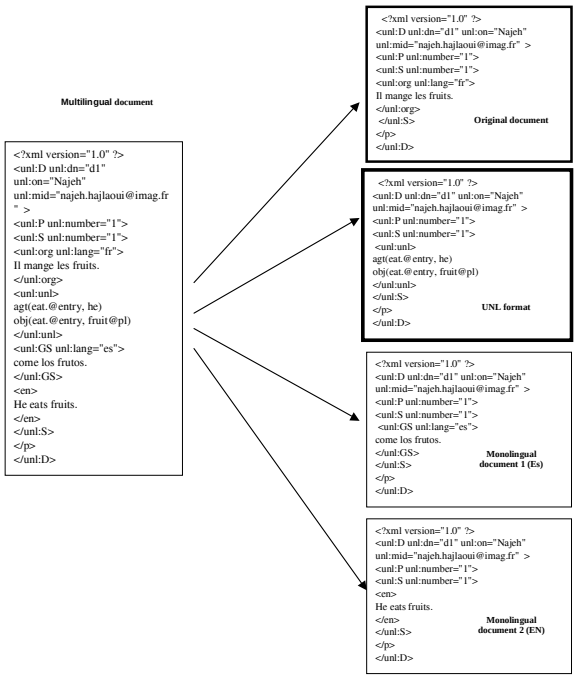


Figure 5: total separation of a multilingual document.

This separation of UNL-xml representation can be improved by gathering technical information common to the monolingual documents in the same document of description. That is possible using XML facilities for creating and managing metadata.

3.3 Discussion

In the first technique of separation

- the autonomous evolution of each linguistic version is possible; that constitutes an important advantage for human revision.
- the source language, the target language and the UNL representation are in the same file, which allows the simple reuse of tools and interfaces of "traditional" MAHT (Machine-Aided Human Translation) systems, there must be a source text and a target text.
- There exist "local" UNL tools which begin to be really used in practice.

In the second technique and since we have only one UNL-xml file, controlled and centralized at the level of sentences, this last file cannot remain strictly parallel with each linguistic version; it has to some extent to reflect modifications. For example, if

we replace in the French file a sentence by two sentences, it will be necessary to leave the UNL graph for the old large sentence in the UNL file and to add 2 new UNL graphs.

Consequently, the two preceding techniques are not satisfactory and there remains the problem of the maintenance of the correspondences.

If modifications are done in all the versions, we cannot use the UNL file as "center" also serving to memorize these modifications.

The principle of our solution is inherited from the area of technical document management and from the PAPILLON project. This solution is based on two important points:

– Monotony: never erase anything in any "volume" (an XML file) but add new evolutionary versions.
– Modularity: represent the correspondences in a separate way.

We propose the following diagram:

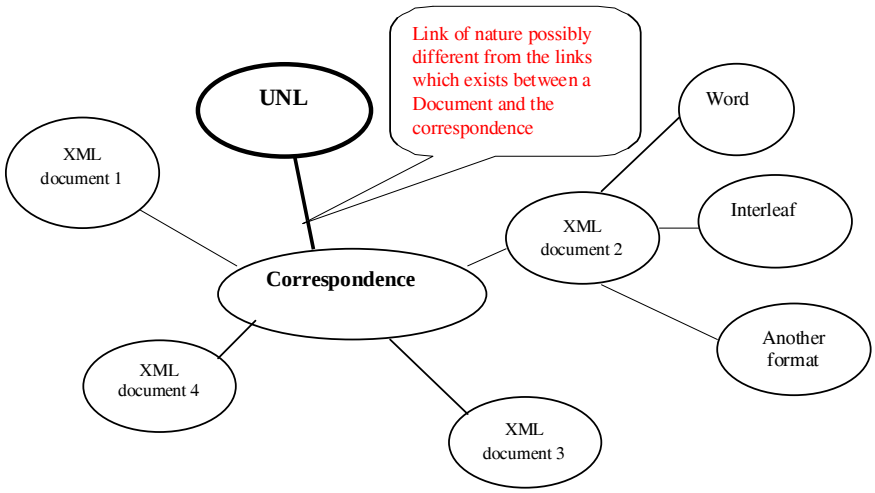


Figure 6: correspondence between several documents.

4 Second solution: a central representation of all correspondences between monolingual and UNL content

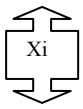
4.1 Logical View

The idea is to represent the correspondences between the various linguistic versions in the form of links in a central structure. These links can be numbers of sentences in the case of a simple local structure such as a large XML file, which includes all the data, the URLs of XML and DTD files representing the versions of each language. It is to some extent a question of following the life cycle of each version, of conserving a complete history of the modifications and applying thereafter the list of the modifications made to the parallel versions to keep alignment.

When a new revision is created, it is necessary to keep a trace identifying the reason for this modification. Moreover, information to be annotated on the object to be replaced in the document is predefined: author, date of operation, optional comment describing the cause of operation.

In what follows, we propose a representation of the correspondence between the linguistic versions which highlights the dependence of the data.

In the figure,



indicates a correspondence link

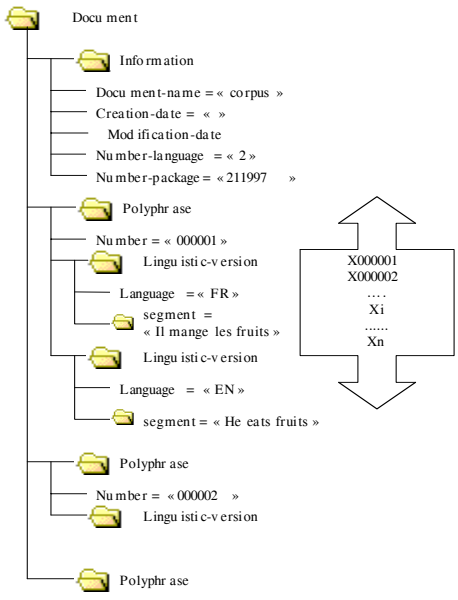


Figure 7: tree XML and representation of correspondences.

The XML tree conforms to the MLD.dtd (Multilingual Language Document). Xn represents a link between the linguistic versions.

For example, X000001 is a link between the French version « Il mange les fruits » and the English version "He eats fruits" constituting the first polyphrase.

We store the set of these links in the XML file, as well as the history of the modifications made to each version.

4.2 Physical view

Data will be stored on a central server in two ways:

- A “Postgres” data base
- File descriptors written in XML, and conforming to a certain DTD, by default our MLD.dtd (Multilingual Language Documents).

The data stored in the database comprises all that relates to the effective management of XML files and access rights on the server. Data tables gather the following information:

- Correspondence between the linguistic versions, their files descriptors in XML, and their DTD.
- URLs of various files (XML, DTD) in order to allow searching and handling them on the server.
- Total information on the level of a version for managing it, and for checking access rights: name, version, author, creation date, planning date in the case of versions under development, translation system used, and some comments.
- Access and modification rights (import and export).
- A version can have two states:
- Private Version : the version is stored on the user workstation, this version can be reloaded and modified.
- Published Version : the version is stored on the server. It results from the decision to publish a Private Version.

4.3 A First Mockup (TraCorpEx project)

After having studied possible architectures and data structures, we have started practical experiments in the framework of the TraCorpEx project. Two parallel corpora in Japanese-English are available [3]. The first comprises 162000 sentences from the CSTAR project and the second 214000 sentences from the PAPILLON project.

To easily manage these corpora using XML, we defined a DTD, MLD.dtd, corresponding to the general structure of multilingual documents. MLD (Multilingual Language Documents) is evolutionary and allows to add other languages to these corpora.

4.3.1 MLD (MultiLingual Documents)

A polyphrase is the set of linguistic versions of the same segment, which have one attribute in common, a unique number. They are also identifiable by other attributes: the language, and for each language the version of the content. In these corpora, the level of alignment is the sentence, but it can go down to a finer level of segments and words. In other corpora, we might go up to the level of paragraph, if sentences are not perfectly aligned.

4.3.2 Interfaces

At this point, the storage format adopted in TraCorpEx is an XML file, which respects MLD.dtd. Upper levels concern the division into corpora, then into sections (import files), then into sentences. Further levels give a hierarchical structure to a polyphrase: language, original and versions, distances, administrative information for tracing etc. At each level, some information is encoded as XML attributes.

```
<!ELEMENT document (information, polyphrase*) >
<!ELEMENT information (#PCDATA) >
<!ATTLIST information document-name CDATA
#REQUIRED>
<!ATTLIST information creation-date CDATA
#IMPLIED>
<!ATTLIST information modification-date CDATA
#IMPLIED>
<!ATTLIST information number-language CDATA
#IMPLIED>
<!ATTLIST information number-polyphrase CDATA
#IMPLIED>

<!ELEMENT polyphrase (linguistic-version*) >
<!ATTLIST polyphrase number CDATA
#REQUIRED>

<!ELEMENT linguistic-version(segment) >
<!ATTLIST linguistic-version language CDATA
#REQUIRED>
<!ELEMENT segment(#PCDATA) >
```

This DTD respects the tree structure of the corpora, as well as the dependencies which arise from the translation process, as we go down the tree towards the contents.
It describes a format for multilingual, multiversion documents with m languages and n versions, n>m, and represents at the same time the correspondences between the parallel parts.
A multilingual document is a set of organizational information (name of the document, creation date, last modification date, numbers of languages, numbers of polyphrase) plus a set of polyphrases.

Figure 8 : MLD (MultiLingual Documents)

To add French to these corpora, we have begun to use the commercial MT system Systran-Pro/EF and to revise the results. We plan to run other MT systems and to choose automatically the "best" translation using the distances between the retrotranslations and the orignal English. In case of conflict, we will also use distances between the translations, to group them, and between translations and original, to detect those with more unknown words, left untranslated.

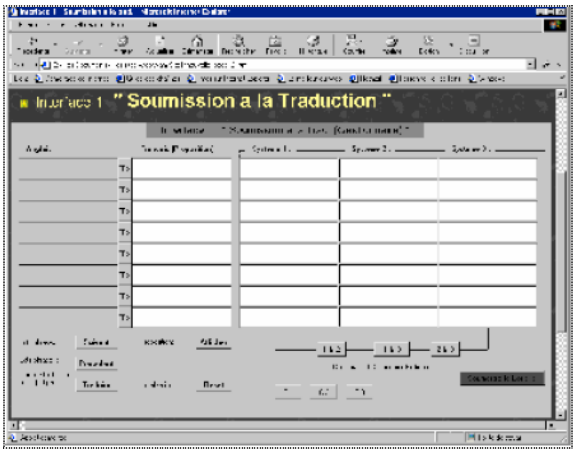
Last but not least, further elaboration, again using string distances, will provide various feedbacks to the developers of the MT systems thus used.

A third interface will be built for the preparation of feedbacks to the developers of the MT systems used. It will allow to calculate and validate the words unknown or badly translated by each system, and to provide translation suggestions from “reference” translations obtained after human revision. It will also provide comparisons between the various systems used, always thanks to the computation of distances at the level of the characters or words.

It also computes distances between English original sentences, so that the document can be used as a translation memory in the following step.

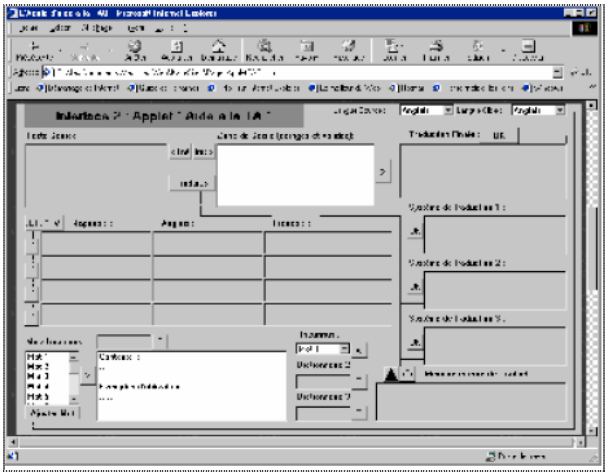
5 Conclusion

The proposed structure of multilingual multiversion documents is technically flexible and modifiable on the initiative of the administrator. It is declared in a hierarchical way in the form of an XML DTD and can be tailored to each corpus of multilingual structured documents aligned at the level of sentences. The hope is that it can contribute to the standardization of multilingual documents, needed to facilitate their management and evolution.



A java program has been developed to calculate the distance between two character strings and to post the result in the form of a matrix and an XML file directly presentable in Word "Track changes" format.
Prototypes of two interfaces have also been produced.
The "preparation" interface allows to submit the English sentences to two or three EF MT systems and to compute the "best" translation of each sentence.

Figure 9 : Interface 1 "preparation"



The second interface is for human revision of the best suggestion using an English zone: we can correct words or expressions and use the translation memory which is in this case the multilingual document itself.

Figure 10 : interface 2 "revision"

The approach presented here is quite flexible and allows any description of file and directory by XML tags, for multiple applications, among which multilingual information retrieval, multilingual summary, multilingual categorization and of course all types of translation.

References

Al-Assimi A.-B. (2000) *Gestion de l'évolution non centralisée de documents parallèles multilingues.* Nouvelle thèse, UJF, Grenoble, 31/10/00, 200 p.

Al-Assimi A.-B. & Boitet C. (2001) *Management of Non-Centralized Evolution of Parallel Multilingual Documents.* Proc. Internationalization Track, 10th International World Wide Web Conference, Hong Kong, May 1-5, 2001, 7 p.

Blanc É. & Sérasset G. (2001) *From Graph to Tree: Processing UNL Graphs using an Existing MT System.* Proc. The first UNL open Conference, Suzhou, China, 22-26 November 2001, UNDL.

Boguslavsky I., Frid N., Iomdin L., Kreidlin L., Sagalova I. & Sizov V. (2000) *Creating a Universal Networking Language Module within an Advanced NLP System.* Proc. COLING-2000, Saarbrücken, 31/7—3/8/2000, ACL & Morgan Kaufmann, H. Uszkoreit ed., vol. 1/2, pp. 83-89.

Boitet C. & Tsai W.-J. (2002) *Coedition to share text revision across languages.* Proc. COLING-02 WS on MT, Taipeh, 1/9/2002, 8 p.

Boitet C. (2003) *Approaches to enlarge bilingual corpora of example sentences to more languages.* Proc. Papillon-03 seminar, Hookaido university, Sapporo, 3-5 July 2003, 13 p.

Hajlaoui N. (2002) *Gestion des versions des composants éléctroniques virtuels.* Rapport de DEA, CSI, INPG, juin 2002, 80 p.

Sérasset G. & Boitet C. (1999) *UNL-French deconversion as transfer & generation from an interlingua with possible quality enhancement through offline human interaction.* Proc. MT Summit VII, Singapore, 13-17 September 1999, Asia Pacific Ass. for MT, J.-I. Tsujii ed., pp. 220—228.

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. COLING-2000, Saarbrücken, 31/7—3/8/2000, ACL, H. Uszkoreit ed., 7 p.

Tomokiyo M., Al-Assimi A.-B. & Boitet C. (2001) *Multilingual documents management by using Universal Networking Language UNL on Alignment Gestion Tool OGA.* Proc. PACLING'01, Fukuoka, 11-14/9/2001, H. Sakaki ed., 7 p.

Tsai W.-J. (2001) *SWIIVRE- a web site for the Initiation, Information, Validation, Research and Experimentation on UNL (Universal Networking Language).* Proc. First UNL Open Conference - Building Global Knowledge with UNL, Suzhou, China, 18-20 Nov. 2001, GETA, CLIPS, IMAG, G. UNDL ed., 8 p.

Emmanuel Planas Ph.D. Thesis. http://bibliotheque.imag.fr/theses/1998/Planas.-Emmanuel/these.dir/

Multilingual corpora (ITC, Italy). http://tcc.itc.it/people/forner/multilingualcorpora.html

UNL project (Universal Networking Langage). http://www.undl.org/

W3-TR/REC, www.w3.org/TR/REC-xml.