PolyphraZ□a tool for the quantitative and subjective evaluation of parallel corpora

Najeh HAJLAOUI, Christian BOITET

GETA, CLIPS, IMAG
Université Joseph Fourier, BP 53
38041 Grenoble, France
{Najeh.Hajlaoui,Christian.Boitet}@imag.fr

Abstract

The PolyphraZ tool is under construction in the framework of the TraCorpEx project (Translation of Corpora of Examples), for the management of parallel multilingual corpora (coding, format, correspondence). It is a software platform allowing the preparation and handling of parallel corpora (languages, codings...), parallel presentation, and addition of new languages to existing corpora by calling several MT systems, and letting human translators produce the final reference translations by using a web-based editor. It integrates the computation of some objective evaluation metrics (NIST, BLUE), and enables subjective evaluations thanks to parallel presentations, and formating based on distance computations between sentences (at several levels). In the future, PolyphraZ should also support versioning and provide feedbacks to developers of the MT systems used: unknown words, badly translated words, and comparative presentations of the outputs of the various systems.

Introduction

We work on several parallel corpora such as the BTEC corpus and the Tanaka corpus, but we miss effective tools for the management of these corpora, such as a web platform allowing import, export, preparation (coding, formats...) and processing (translation, revision, edition…) of multilingual corpora.

The BTEC comprises about 162320 sentences (about 4000 standard "translator's pages"[1]) in Japanese, Chinese, English and Korean, less in other CSTAR languages. Diffusion of this corpus is restricted to ATR partners in CSTAR (consortium for speech translation advanced research). Our practical goal is to produce a French version of the BTEC, with the same quality of the sentences.

Tools such as Excel, TextEdit or BBEdit do not allow sharing such corpora on the Web, nor editing and visualizing parallel sentences.

During a stay at ATR, the second author translated the complete BTEC, submitting 163 files of 1000 sentences to Systran Premium v.4, adequately parametrized, and revised the first 1000 sentences, equivalent to 24 standard translator pages in 6:08 hours, or 15 mn per page, under TextEdit, a standard text editor, manually aligning the source and target files. In a later experiment, he did the same on 510 sentences while three other French native speakers translated them by hand, at a rate of 1h per page each (the usual figure in professional translation). That shows that using MT output really speeds up the process of producing (good) reference translations in a new language, but that sharing the workload is still a necessity (it is about 5000 hours for the whole BTEC with no machine help, and still about 1000 hours using MT outputs as suggestions[2]).

He also tried to use Excel on a larger batch of sentences (20000 sentences or 480 pages), but, as shown in figure 1, the gain was quite small, although alignment is automatic. The reason is that saving takes too much time on such a large file.

Corpus	Num_fich	Num_deb	Num_fin	Pareg	Mn_rev	Cumul	Pages std	Mn/page
CSTAR	BTEC	000001	163000					
		1	1000	1000	368,1	368	24,06	15,30
translation:	Systran v4	1001	2000	1000	300	668	24,06	12,47
		2001	2200	200	55	723	4,81	11,43
revision:	*TextEdit*	2201	2286	86	30	753	2,07	14,50
	Total				753,10		54,99	13,69
					12:33			0:14
IWSLT-04	TRAINING	JE00001	JE20000					
		1	100	100	25	25	2,41	10,39
		101	166	66	20	45	1,59	12,60
revision:	*Excel*	167	301	135	42	87	3,25	12,93
		302	374	73	23	110	1,76	13,10
		375	602	228	80	190	5,48	14,59
		603						
					190,00		14,48	13,12
					3:10			0:13

Figure 1 : revision times with TextEdit and Excel

There is some dilemna if translation is performed on parts of a corpus: on one side, files should be large so that global changes, which are very frequent and productive, can concern as many sentences as possible, and on the other side, files should be small, so that each can be translated in a reasonably short time.

Our goal, then, is to develop an efficient tool to expand a multilingual corpus into other languages, as a whole, but in a distributed way, by the cooperation of several translators cooperating through the Web.

We will call such a corpus a "polyphrase" memory (MPM), introducing the term "polyphrase" instead of "sentence" because there may be several "proposals" (or "paraphrases") in each language for one "original" sentence in one language.

In late 2003, we have started work on PolyphraZ, a web server for displaying, translating and editing MPMs on the Web. Last July, the first 2 functionalities had been implemented, and they have been used experimentally in the context of the CSTAR MT evaluation campaign.

We now present in more detail the context and the objectives of the TraCorpex project and the PolyphraZ tool. We then describe the architecture of PolyphraZ, as well as intended scenarios of use and types of users.

[1] 250 words, 1400 characters, corresponding to A4, Times 12, double space.

[2] Note that they cannot be used alone, by a monolingual posteditor, but they must be shown with the original sentences.

1. TraCorpEx

1.1. Context

We started the TraCorpEx project because we realized that very similar tasks had to be undertaken in 3 other projects: the Papillon project (Papillon) of cooperative construction of a large multilingual lexical data base on the Web, the C-STAR III project of translation of spoken dialogues, a French and Tunisian project (Hajlaoui, Boitet, 2003b), the UNL project (UNL) of communication and multilingual information system, and some PhD projects.

1.2. Current situation

To get our hands on concrete data, we initially concentrate on 2 parallel corpora, structured differently.

- The BTEC corpus created by the C-STAR project is made of sets of 163 files of 1000 sentences, one seet per language. It is complete in English, Japanese), Chinese and Korean. File size ranges from 12K to 40K, for a total of 6.1 Mo for each language. Several coding systems are used (8859-1 for English, EUC-JP for Japanese, etc.).

- The TANAKA corpus (Japanese-English), given to the Papillon project a few months before the death of its author in 2002, is made of 45 bilingual files for a total of 18.4 Mo. It contains sentences of newspapers or teaching material published by NHK for Japanese learning English.

Later, we will consider corpora from the UNL project, where each document is a multilingual file containing for each sentence its text in the source language, a UNL graph, the result of deconversions in a certain number of languages, and possibly their revisions, or direct manual translations.

All these parallel corpora are aligned at the level of sentences. It is interesting to make it possible to go to a finer level like the segments and the words. In other corpora, we will be obliged to go up to the level of paragraphs, because the sentences are not aligned perfectly.

The first problem raised by the parallel available corpora it is that there is no tool making it possible to visualize their contents at a glance, sentence by sentence, nor to show the fine correspondences between subsentential segments. In addition, in the case of UNL documents, we cannot visualize at the same time an utterance in several languages and the corresponding UNL graph. Lastly, it is never possible to see the successive versions at the same time.

1.3. Detailed objectives

The objectives of TraCorpEx project are as follows.

1.3.1. Addition of new languages (horizontal expansion)

Starting from parallel corpora, we want to add one or more languages (those of the Papillon project for the Tanaka corpus, French and Arabic for the BTEC corpus). The final results must be of high quality, to be usable as "reference translations". Hence, humans must participate in the translation work, and we have to develop some kind of translation aids (TA), sharable on the web.

1.3.2. Building a software platform for translation

A subgoal, then, is to develop a web-enabled platform to import corpora and put them in some normalized form, to translate them using various translation aids (multilingual editor, translation memories, dictionaries, and distant MT systems), to visualize and evaluate them, and to export the results in various formats and codings. To encourage MT developers to give free access to some versions of their products, it is also needed to offer various kinds of feedbacks to developers. That is the PolyphraZ project.

1.3.3. Enlarging parallel corpora (vertical expansion)

A third goal of TraCorpEx is to research and implement techniques to enlarge an MPM by creating neew polyphrases. Interesting results have already been obtained by Y. Lepage, using a combination of analogical computing and n-gram filtering.

2. PolyphraZ

2.1. Additional goal: evaluation

PolyphraZ should also make it possible to evaluate automatic translators with automatic methods such as NIST, BLUE, PER, and to use this possibility in CSTAR, to evaluate the Chinese-English and Japanese-English translations. To evaluate the results of various MT systems will also enable us to determine "the best" (or less bad!) translation, proposable to a contributor as a starting point for revision.

The quality of the translations should also be evaluated using calculations of distances between sentences and reverse translations.

2.2. Feedbacks to developers of MT systems

We also want to give feedbacks to the developers of the systems used (unknown words, badly translated sentences...), and a comparative presentation between the various translation systems.

The whole of the objectives of this project led us to propose interactive Web interfaces allowing us to chooses, use, compare, publish machine translations corresponding to several language pairs, and to contribute to the improvement of the results by sending feedbacks to the developers of these systems.

2.3. General architecture

We follow the software architecture of the Papillon platform, and reuse certain techniques of parallel visualization of translation memories (PhD thesis of Ch. Chenon).

We classify the objects to handle in three types:

- Raw corpus sources

- Sources transformed into our CXM (Common Example Markup) XML format (coded in UTF-8), for visualization "just as they are", and then in the CPXM format for parallel visualization.

- MPM: multilingual polyphrase memory

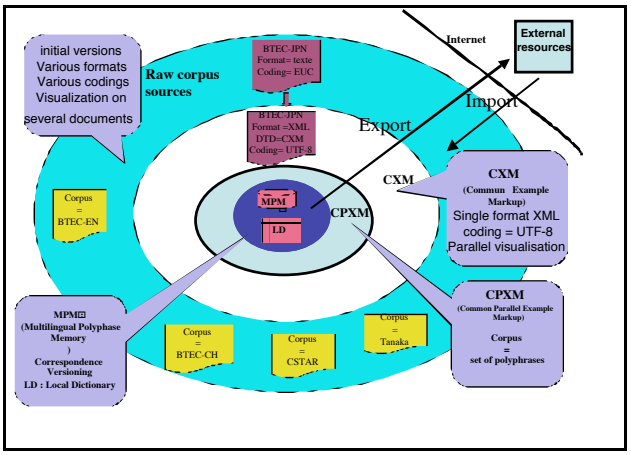

Figure 2: objects of the PolyphraZ platform

2.4. Users of PolyphraZ

We distinguish four principal users: the preparer, the reader ("normal" user), the reviser and the manager.

2.4.1. The preparer

His role consists in calling translation systems, by parameterizing them as well as possible, which supposes a certain linguistic ability, and can require a delicate development: comparison of the results obtained with various parameter settings, segmentation in "blocks" corresponding to various "optimal" parameter settings, etc.

The preparer can launch automatic evaluations (NIST, BLUE...) on results of translation, and the computation of distances between sentences (results of translation and/or reverse translations). The mixed character and word distance computation produces, in addition to a value, an XML string from which a “track changes” presentation can be generated. He can also set the parameters determining "the best" suggestion among various translation candidates.

2.4.2. The reader (normal user)

A reader can visualize the data (the original, various translations, and distances between the character strings), but is not allowed to edit the translations.

2.4.3. The translator-posteditor

The translator-posteditor is a contributor who translates from scratch or revises proposed translations (in general, MT results or sentences retrieved from translation memories). There is an editable area to modify the active sentence. One can also ask for global modifications (ex: "SVP" changed into "s'il vous plait" in transcribed spoken utterances) and correct or supplement the local dictionary attached to the MPM. The system uses the reference sentences already produced like a translation memory. PolyphraZ will also be a system of assistance to the translator, limited to the translation of sets of sentences (or titles).

2.4.4. The manager

The last type of user is the manager, who will request from PolyphraZ "feedbacks" for the developers of the MT systems used. A manager cans itself be a developer of an MT system. He can draw up (thanks to calculation of distances and to an adapted presentation) a list of unknown words and words badly translated by each system, then validate it, propose for these words suggestions of translation from the "reference" translations obtained after human revision and provide a presentation of the evaluations and comparisons between the results of the various systems used or their various parameter settings.

2.5. Implementation

PolyphraZ is multi-platform (Mac OS-X, Unix, Linux Windows), being programmed in standard java under the Enhydra development environment used for the dynamic and multilingual Papillon web site.

3. Scenarios of use

The following diagram synthesizes possible uses.

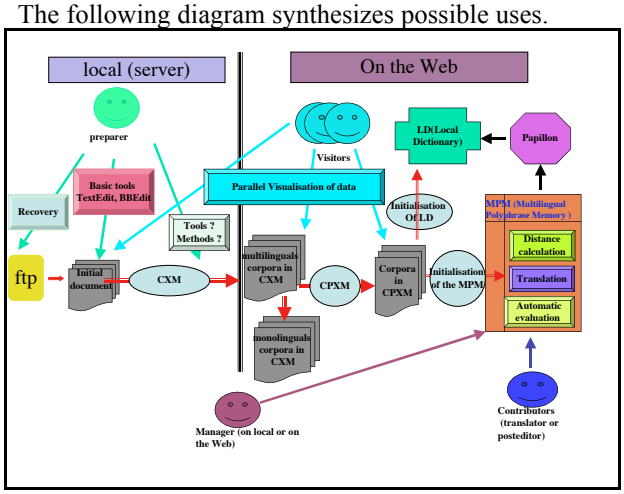

Figure 3 : scenarios for using PolyphraZ

3.1. Import into CXM (Common eXample Markup)

```
<?xml version="1.0" standalone="no" ?>
<!DOCTYPE document SYSTEM "CSTAR_BTEC_DTD.dtd" >
<document>
 <information documentname="CSTAR-corpus BTEC EJ"
 creation-date="Tue May 21 JST 2002"
 modification-date="Tue May 21 JST 2002"
 coding-set="UTF-8"
 number-of-language="2"
 number-of-sentences="162320"/>
 <sentence sentence-id="000001">
 <sentence xml:lang="EN">
 <segment segment-id="1">
 Hamburger and stew on the right side and salad, please.
 </segement>
 </sentence>
 <sentence sentence-id="000001">
 <sentence xml:lang="IT">
 <segment segment-id="1">
 Humburger e stufato dalla parte destra e insalata,per favore.
 </segement>
 </sentence>
 </document>
```

Figure 4: example XML file conforming to the CXM.dtd

When we import a corpus, we transform it in a single coding (UTF-8), and a single XML format, CXM, similar to the CDM (Common Dictionary Markup) format of the Papillon project.

3.2. CPXM.dtd (Common Parallel eXample Markup)

A second consists in transforming all CXM files corresponding to a given multilingual parallel corpus

into a file in the CPXM format (see appendix 2). In this format, we introduce the <polyphrase> XML element, which contains a set of monolingual components, each component containing possibly one or more proposals.

3.3. MPM.dtd (Multilingual Polyphrase Memory)

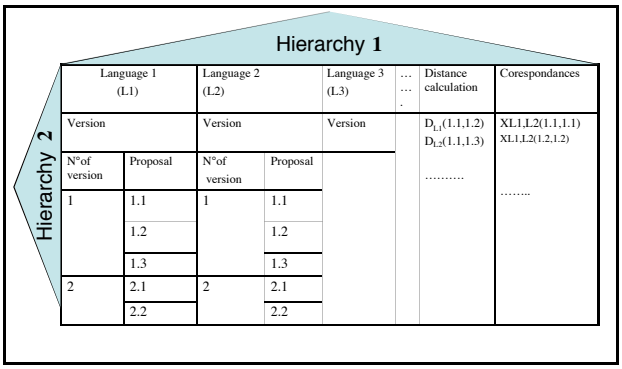

Figure 5 : logical view of a MPM

The MPM format is still undergoing changes. The current version is given in appendix 3.

The current version of PolyphraZ is not complete, but several functionalities are already usable, and accessible on the [TraCorpEx] website.

3.4. Parallel visualization

PolyphraZ proposes an option common to the three preceding stages, which consists in visualizing in a parallel way the "columns" of the polyphrases, to allow for manual (subjective) comparison of the translations. It is actually useful for readers, translators revisors, and managers.

Figure 6 shows an example taken from the BTEC corpus. At this moment, the width of the columns is fixed, but it should be controllable by the user in the future, as well as the display of evaluations and distance computations.

Phrase Identifier	English	Chine	Korea
000001	Hamburger and stew on the right side and salad, please.	请来个汉堡，右边的那个炖菜和色拉。	햄버거랑 오른쪽에 있는 스튜랑 샐러드로 할게요.
000001	****	****	햄버거하고 오른쪽에 있는 스튜하고 샐러드 주세요.
000002	That fried fish, one sausage with green peas, please.	请来个那种炸鱼，一个加豌豆的香肠。	저 생선 튀김하고 완두콩 소시지 주세요.
000002	****	****	프라이드 피시하고 소시지에 완두콩 얹어 주세요.
000003	T-bone steak and sauerkraut and fried potatoes, please.	请来个T形的牛排和德国泡菜还有炸薯条。	티본 스테이크랑 사우어크라우트, 프라이드 포테이토로 할게요.
000003	****	****	티본 스테이크하고 양배추 절임하고 감자 튀김 주세요.
000004	Roast chicken and two slices of ham on this side and spinach, please.	请来个烤鸡肉和两片靠这边的火腿还有菠菜。	닭고기 구이하고 햄 두 조각하고 시금치로 할게요.

Figure 6: parallel visualisation of the BTEC

3.5. Evaluation of translation results

We have programmed and integrated in PolyphraZ three evaluation methods (NIST, BLUE, and distance computation). NIST and BLUE are well known. As far as distances are concerned, we use a combination between two edit distances, one based on characters and the other on words.

The edit distance between two strings of atoms (characters or words) is the minimal number of suppressions, insertions or replacements of atoms

necessary to transform one string into the other. In each case, the set of atoms is the union of the atoms in the 2 strings.

The cost of inserting, suppressing and exchanging characters is defined beforehand by a table or by 3 functions. The cost of inserting or suppressing a word is its character distance with the empty word, and the cost of exchanging 2 words is their character-based edit distance.

At any level, the edit distance between two strings $\mathbf{x} = a_1 \ldots a_m$ and $\mathbf{y} = b_1 \ldots b_n$ is D(m, n), defined by:

$$\begin{cases} D(0,0) = 0 \\ D(0,1) = C(\square, b_1) \\ D(1,0) = C(a_1, \square) \\ D(i+1, j+1) = \min \begin{cases} D(i+1, j) + C(INS(b_{j+1})), \\ D(i, j+1) + C(DES(a_{i+1})), \\ D(i, j) + C(SUB(a_{i+1}, b_{j+1})) \end{cases} \end{cases}$$

The mixed distance is then defined by:

$D = \alpha\, D_{char} + (1-\alpha)\, D_{word}$; $0 \leq \alpha \leq 1$.

For the moment, we use the well-known dynamic programming algorithm of Wagner and Fischer (Wagner & Fischer, 1974), but it will be easy to replace it by more efficient ones in the future.

3.6. Interfaces

Prototypes of two interfaces have been produced.

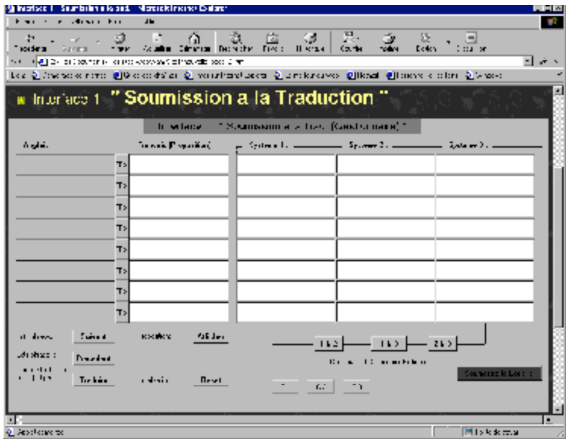

Figure 7 : Interface 1 “preparation”

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

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.

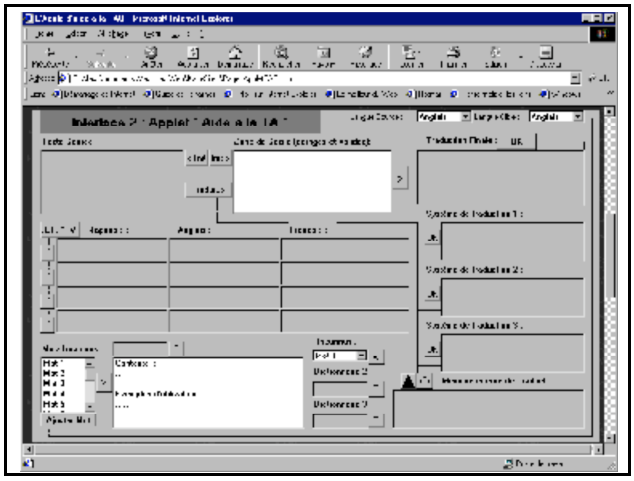

Figure 8 : interface 2 “revision”

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.

Conclusion

The external and middle levels of PolyphraZ are already used. They allow us to put imported multilingual corpora of parallel sentences into a common format and encoding (CXM), and then to transform a whole corpus into one or more files in CPXM formats, and visualize their content on the web. The central level of MPM (Multilingual Polyphrase Memory) is almost completed. It will also support versioning.

In the future, we plan the MPM form to use not only to extend to new languages, but also like a "pivot", to establish the correspondence between monolingual structured documents corresponding. To each other even if they are not perfect and complete mutual translations or, if they are complete mutual translations, without imposing a strict alignment of sentences, paragraphs, sections, etc.

We are also studying how to integrate into a MPM structure "generators" specifying a class of sentences (automata for messages with variables and variants, regular expressions for the IF of CSTAR, etc), and to use them to extend a MPM "in width" (addition of new languages), but also "in height", by the automatic creation of new statements, natural and/or formal.

References

[1] A. Assimi (Assimi, 2000). *Gestion de l'évolution non centralise de documents parallèles multilingues*, Nouvelle these, UJF, Grenoble, 31/10/00, 200 p.

[2] A-B. Assimi & C.Boitet (Assimi & Boitet, 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.

[3] Ch. Boitet (Boitet, 2003) *Approaches to enlarge bilingual corpora of example sentences to more languages*, Papillon-03 seminar, Sapporo, 3-5 July 2003, 12 p.

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

[5] H. Vo-trung (Vo-trung, 2004) *Réutilisation de traducteurs gratuits pour développer des systèmes multilingues*, RECITAL 2004, avril 2004, Fès, Maroc.

[6] N. Hajlaoui, Ch. Boitet (Hajlaoui & Boitet, 2003a), *A "pivot" XML-based architecture for multilingual, multiversion documents : parallel monolingual documents aligned through a central correspondence descriptor and possible use of UNL*, Convergences'03, Alexandria, 2-6 December 2003.

[7] N. Hajlaoui, Ch. Boitet (Hajlaoui & Boitet, 2003b), *Modélisation de la production de phrases, projet franco-tunisien entre l'équipe* GETA, CLIPS, UJF, Grenoble et université de Sousse, Tunisie, 25 p.

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

[9] R. Wagner & M. Fischer (Wagner & Fischer, 1974) *The String-to-String Correction Problem* ACM Journal of the Association for Computing Machinery, Vol. 21, No 1, Janvier 1974.

[10] W.-J.Tsai (Tsai, 2001) SWIIVRE *a web site for the Initiation, Information, Validation, Research and Experimentation on UNL*. Proc. First UNL Open Conference - Building Global Knowledge with UNL, Suzhou, China, 18-20 Nov. 2001, 8 p.

[11] (C-STAR-III) *Projet C-STAR*, http://www.c-star.org/

[12] (Papillon) *Projet PAPILLON de construction coopérative d'une base lexicale multilingue et de construction de dictionnaires,* http://www.papillon-dictionary.org/

[13] (TraCorpEx) projet TraCorpEx http://www-clips.imag.fr/geta/User/najeh.hajlaoui/tracorpex/index.html

[14] (UNL) *projet UNL (Universal Networking Langage)*, http://www.undl.org/

Appendices

```
<!-- CXM.dtd  (Commun eXample Markup ) is a
DTD  which  describes
the corpora (multilingual or monolingual),
it is the simplest format for the recovered
data.
$Author:  Najeh Hajlaoui
najeh.hajlaoui@imag.fr
 $Date: 2004/07/06 10:28:30 $
 -->
<!ELEMENT document (information, sentence*)
>
<!ELEMENT information (#PCDATA) >
<!ATTLIST information   document-name
CDATA #REQUIRED>
<!ATTLIST information   creation-date
CDATA #IMPLIED>
<!ATTLIST information   modification-date
CDATA #IMPLIED>
<!ATTLIST information   coding-set    CDATA
#IMPLIED>
<!ATTLIST information   number-of-languages
CDATA #IMPLIED>
<!ATTLIST information   number-of-sentences
CDATA #IMPLIED>


<!ATTLIST sentence   sentence-id    CDATA
#REQUIRED>
<!ELEMENT sentence (language*) >

<!ATTLIST language xml:lang CDATA #REQUIRED>
<!ELEMENT language (segment*) >
<!ATTLIST segment   segment-id    CDATA
#REQUIRED>
<!ELEMENT segment (#PCDATA) >


<!-- Document is a set of sentences, each
sentence is defined
by an identifier called sentence-id. each
sentence is e set of language  -->
<!-- number-of-languages is the total number
of languages constituting
the document; if the document is
monolingual, number-of-languages =1   -->
<!-- number-of-sentences is the total number
of sentences
constituting the document -->
<!-- Each language is a set of one or more
possible  segment; each segment is
identified by an attribute called segment-id
-->

<!-- Each sentence is a set of one or more
possible  segment; each segment is
identified by an attribute called  segment-
id -->
```

Appendix 1 : CXM.dtd (Common eXample Markup)

```
<!-- CPXM.dtd  (Commun Parallel eXample Markup
) is a  DTD  which  describe
the multilingual documents (m languages),
multiversions(n versions) (n>m),
it allows the description of the notion of the
ploy phrase represented by a single coding.
In CPXM, we  can introduce a new proposals.
 $Author:  Najeh Hajlaoui
najeh.hajlaoui@imag.fr
 $Date: 2004/07/06 15:28:30 $
 -->
<!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   coding-set    CDATA
#IMPLIED>
<!ATTLIST information   number-of-languages
CDATA #IMPLIED>
<!ATTLIST information   number-of-polyphrases
CDATA #IMPLIED>

<!ATTLIST polyphrase   polyphrase-id    CDATA
#REQUIRED>
<!ELEMENT polyphrase (monolingual-component*) >

<!ATTLIST monolingual-component xml:lang CDATA
#REQUIRED>
<!ELEMENT  monolingual-component (proposal*) >
<!ATTLIST proposal   proposal-id    CDATA
#REQUIRED>
<!ELEMENT  proposal (#PCDATA) >

<!-- number-of-languages is the total number of
languages appearing in the document;
if the document is monolingual, number-of-
languages =1   -->
<!-- number-of-polyphrases is the total number
of polyphrases constituting the document -->
<!-- A polyphrase is a set of monolingual
components, each containing 1 or more possible
proposals.
 Every polyphrase is identified by a number
called polyphrase-id -->
<!-- Each monolingual component is a set of one
or more possible renderings of the proposal in
question;
it is identified by an attribute which
indicates the language -->
<!-- Proposal represents the level of
alignment, it is usually a sentence -->
```

Appendix 2 : CPXM.dtd

```
<!-- MPM.dtd  (Multilingual Polyphrases
Memory ) is a  DTD  which  allows the
generation of sentences aligned in several
languages
and the management of the corespondance
between these sentences.
$Author:  Najeh Hajlaoui
najeh.hajlaoui@imag.fr
$Date: 2003/01/28 21:28:30 $ -->
<!ELEMENT document (information, generator*,
node-of-correspondence*) >
<!ELEMENT information (#PCDATA) >
<!ATTLIST information document-name  CDATA
#REQUIRED>
<!ATTLIST information   creation-date
CDATA #IMPLIED>
<!ATTLIST information modification-date
CDATA #IMPLIED>
<!ATTLIST information   coding-set    CDATA
#IMPLIED>
<!ATTLIST information number-of-languages
CDATA #IMPLIED>
<!ATTLIST information number-of-generator
CDATA #IMPLIED>

<!ELEMENT generator (instance*) >
<!ATTLIST generator   original    CDATA
#REQUIRED>
<!ATTLIST generator   context    CDATA
#REQUIRED>

<!ELEMENT  instance (segment*) >
<!ATTLIST instance xml:lang CDATA #REQUIRED>
<!ATTLIST segment node-of-correspondance-id
CDATA #REQUIRED>
<!ELEMENT  segment (proposal) >
<!ELEMENT  proposal (#PCDATA) >

<!-- number-of-languages is the total number
of languages appearing in the document; if
the document is monolingual, number-of-
languages =1   -->
<!-- number-of-generator is the total number
of generator appearing in the document -->
<!-- A generator is a set of original
sentences and  their instance  -->
<!-- A instance is a set of one or more
possible renderings of the segment in
question; it is identified by an attribute
which indicates the language -->
<!-- Segment represents the level of
alignment, it is usually a sentence -->
<!-- A node-of-correspondence-id represent
the link of corespondance between the
diférents proposals of translation -->
```

Appendix 3 : MPM.dtd