PolyphraZ : a tool for the management of parallel corpora

Najeh HAJLAOUI
GETA, CLIPS, IMAG
Université Joseph Fourier, BP 53
38041 Grenoble, France
Najeh.Hajlaoui@imag.fr

Christian BOITET
GETA, CLIPS, IMAG
Université Joseph Fourier, BP 53
38041 Grenoble, France
Christian.Boitet@imag.fr

Abstract

The PolyphraZ tool is being developed in the framework of the TraCorpEx project (Translation of Corpora of Examples), to manage parallel multilingual corpora through the web. Corpus files (monolingual or multilingual) are firstly converted to a standard coding (CXM.dtd, UTF8). Then, they are assembled (CPXM.dtd) to visualize them in parallel through the web. In a third stage, they are put in a Multilingual Polyphraz Memory (MPM). A "polyphrase" is a structure containing an original sentence and various proposals of equivalent sentences, in the same and other languages. An MPM stores one or more corpora of polyphrazes. The MPM part of PolyphraZ has 3 main web interfaces. One is a web-oriented translator workstation (TWS), where suggestions or translations come from the MPM itself, which functions as its own translation memory, and from calls to MT systems. Another serves to send sentences to MT systems with appropriate parameters, and to run various evaluation measures (NIST, BLEU, and distance computations) in order to propose to the translator a "best" proposal. A third interface is planned for giving feedbacks to the developers of the MT systems, in the form of lists of unknown or wrongly translated words, with suggestions for correct translations, and of parallel presentation of pairs of translations showing the "editing work" to be done to get one from the other. The first 2 stages are operational, and used for experimentation and MT evaluation on the CSTAR 5-lingual BTEC corpus and on the Japanese-English Tanaka corpus used as a source of examples in electronic dictionaries (JDict, Papillon). A main goal of this effort is to offer occasional and volunteer translators and posteditors access to a free TWS and to sharable translation memories put in the MPM format.

1 Introduction

Due to Internet grow, the number of available documents grows dramatically. There is a strategic need for companies to produce and manage information written in more than 30 languages (HP, IBM, MS, Caterpillar). This requires powerful tools to manage multilingual documents.

Current techniques for handling multilingual documents use large-grained linking (at the level of HTML pages), but don't allow fine-grained synchronization (at paragraph or sentence level) and don't permit bilingual or multilingual editing through the Web.

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

- make it possible to use Machine Aided Human Translation (MAHT) techniques, in particular translation memories, for translating and postediting multilingual documents.
- add UNL tags at sentence level to store the translations as well as UNL hypergraphs (anglosemantic interlingual representations), from which raw (or rough!) translations into other languages can be obtained from distant "deconversion" servers.

Here, we are not concerned with the problem of aligning parallel monolingual documents, or realigning them after they have been modified, a frequent need in the case of leaflets and booklets. (Assimi,2000) proposed a tool to handle the non-centralized management of the evolution of multilingual parallel documents. We consider the case, frequent in the industry, where documents are managed centrally, even if they are distributed on several sites. What happens in general is that they are aligned at the level of large blocks, with one file per block and language (fileXXX.en.htm, fileXXX.fr.htm etc. for HTML pages).

What we propose is to align them at the level of sentences, but of course not to have one file per sentence. Rather, if there are N languages, for a given "block" corresponding to some unit of processing (e.g. visualization), we will have either

N monolingual sentence-aligned files, or 1 multilingual file. In both cases, sentences or place holders for sentences will be linked to a MPM to manage translation and postedition.

We began to build PolyphraZ in the context of the TraCorpEx project (Translation of Corpora of Examples). A more recent motivation is to extend the BTEC corpus of CSTAR III (163000 sentences in tourism) to French and Arabic, and to evaluate various Chinese-English MT systems on it.

We will first present the data we start with, and our goals in more detail. In a second part, we will describe the architecture of PolyphraZ, starting from scenarios of use and types of users. Lastly, we will describe the current status of this work.

2 TraCorpEx and PolyphraZ

2.1 Context

The TraCorpEx project has several contexts: the Papillon project (Papillon) of co-operative construction of a large multilingual lexical base on the Web, the C-STAR III project (C-STAR III) of translation of spoken dialogues, a French and Tunisian project (Hajlaoui, Boitet, 2003b), the UNL project (UNL) of communication and multilingual information system, and the PhD research of the various participants in this project.

2.2 Current data and problems

We have initially 2 "parallel" corpora, structured differently.

- The BTEC corpus of C-STAR is made of 5 sets of 163 files of 12K to 40K, each containing 1000 sentences, in English, Japanese (coded in EUC), Chinese and Korean, for a total of 6.1 Mo per language.
- 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 files for a total of 18.4 Mo. It contains sentences of newspapers or teaching works of NHK for the training of English by the Japanese. Each file is bilingual.

We have also corpora from the UNL project, where each document is a multilingual file containing for each sentence its text in 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. As it would be interesting to show correspondences at finer levels (syntagms, chunks, words), we design PolyphraZ to later add tools for subsential alignement such as the one developed by Ch. Chenon for his Ph.D.

In other corpora, we may be obliged to go up to the level of paragraphs, because sentences will not be aligned perfectly. That will not be done completely in PolyphraZ, but at the level of the structure of the multilingual document itself: if 2 sentences are translated by 3, each of the 5 sentences will be in a different polyphrase, with their individual translations, and there will be another polyphrase, of "n-m" type, to contain the 2 complete segements.

The first problem we encounter with the available parallel corpora it is that there is no tool 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 a sentences in several languages and its corresponding UNL graph. Lastly, it is not possible to see successive versions in parallel.

When it comes to evaluation, we can only see the monolingual files, and associated statistical measurements (NIST, BLEU...), but we can never confront them with the real translations and make a direct subjective evaluation.

2.3 Detailed objectives

The objectives of TraCorpEx project are as follows.

2.3.1 Construction of a software platform

We want to build an environment, which supports the import and the export of parallel corpora, the preparation of the data for automatic translators, the postedition (HAMT), the evaluation (various feedbacks methods) and finally a preparation of "feedbacks" to the developers of used MT systems.

2.3.2 Addition of new languages

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).

2.3.3 Evaluation of MT systems

We also wish that the same platform makes it possible to evaluate automatic translators with automatic methods such as NIST, BLEU, 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.

We also want to test a hypothesis by the second author: the quality of the translations could also be evaluated using calculations of distances between sentences and reverse translations.

2.3.4 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.

2.4 The PolyphraZ platform

PolyphraZ is a software platform making it possible at the same time to visualize the available corpora on the Web by showing several languages, with the choice of the user and to work on a basis of "polyphrases" initialised from these corpora while making it possible to control all functions described above (call of MT systems, distance computation, collaborative postedition, evaluation).

2.4.1 General architecture

We follow the software architecture of the Papillon platform.

We classify the objects to handle in three types

- Raw corpus sources
- Sources transformed into our XML format CXM. (Common Example Markup) and coded in UTF-8, for visualization "just as they are", then in CPXM format, DTD for parallel visualization.
- MPM: multilingual polyphrase memory

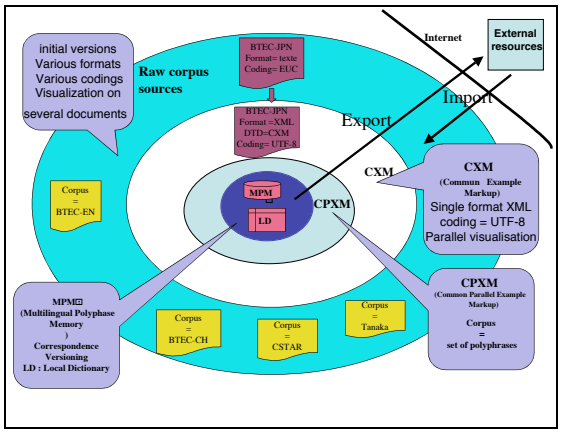

Figure 1: objects of the PolyphraZ platform

2.4.2 Intended users of PolyphraZ

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

- **The preparer**

  His role consists in calling translation systems, thereby parameterizing them as well as possible, which supposes a certain linguistic ability (to compare the results of various parameter settings, and of various segmentations in "blocks", each corresponding to some parameter settings).

  The preparer can also call objective evaluation methods (NIST, BLEU...) on the results of translation, tune with parameters to compute distances between sentences (results of translation and/or reverse translations), and post the results. The distance computation produces, in addition to a value, a XML string from which a “track changes” presentation can be generated. The preparer can also set the parameters determining "the best" suggestion among the various translation candidates.

- **The reader (normal user)**

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

- **The translator-posteditor**

  The translator-posteditor is a contributor who translates from scratch or revises proposed translations (MT results or translations of similar sentences found in the MPM or in other TM put in CPXM or MPM format). 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 is thus also a system of assistance to the translator, limited to the translation of sets of sentences (or titles), with less functionalities than commercial TWS, but usable for collaborative volunteer work by non-professionals.

- **The manager**

  The last type of user is the manager, who will produce from a MPM "feedbacks" for the developers of the MT systems used. A manager cans himself be a developer of an MT system.

  He can draw up a list of unknown words and words badly translated by each system (produced from the traces of distance computations). A second function is to propose for these words suggestions of translation from the "reference" translations obtained after human

revision. Finally, it is possibile to provide a presentation of the evaluations and comparisons between the results of the various systems used and/or their various parameter settings.

2.4.3 Implementation of PolyphraZ

Programmed in standard Java under the Enhydra development environment used for the dynamic and multilingual Papillon web site, PolyphraZ is multi-platform (MacOS-X/Unix/Linux, Windows).

2.5 Scenarios

The use of PolyphraZ can be divided in 3 parts: setting of the data under three different formats (CXM, CPXM, MPM).

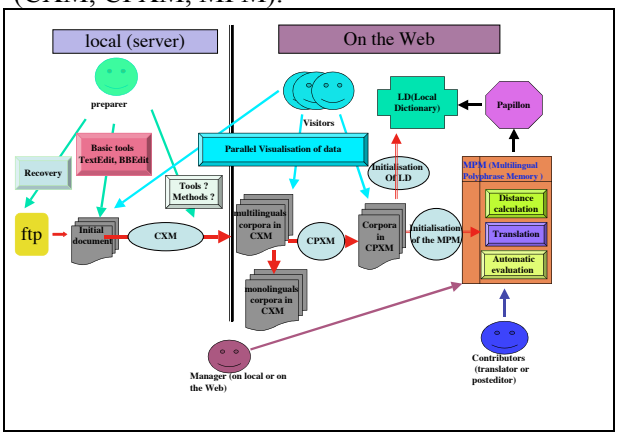

Figure 2 : scenarios for using PolyphraZ

2.5.1 CXM (Common eXample Markup)

In order to manipulate a single format (XML) and a single encoding (UTF-8), we automatically convert into the CXM format the imported data (corpus, text aligned...). CDM is defined in the same spirit as the CDM (Common Dictionnary Markup) of the Papillon project.

```
<?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 3: example XML file conforming to the CXM.dtd

2.5.2 CPXM.dtd (Common Parallel eXample Markup)

A second Java program transforms all CXM files corresponding to a given multilingual parallel corpus of sentences to the CPXM format (see appendix 2). In this format, we introduce the "polyphrase" XML element, which is a set of monolingual components, each containing possibly one or more proposals.

2.5.3 MPM.dtd (Multilingual Polyphrase Memory)

The MPM data structure is under construction. It is intended for the management of the correspondences between the various linguistic versions as well as the modifications which can be made, and to keep the history of the modified files. As shown in the following figure, a MPM of PolyphraZ can contain a set of versions and alternatives of the sentences, as well as the results of various computations.

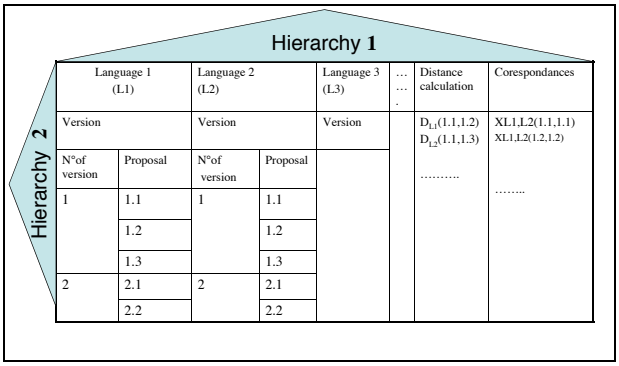

Figure 4 : logical view of a MPM

We give a first version of the MPM DTD in appendix 3.

2.5.4 Parallel visualization

PolyphraZ can visualize polyphrases in parallel from corpora in CPXM or MPM formats. This functionality is useful to compare translations, and is made available to readers; translators revisors, and managers.

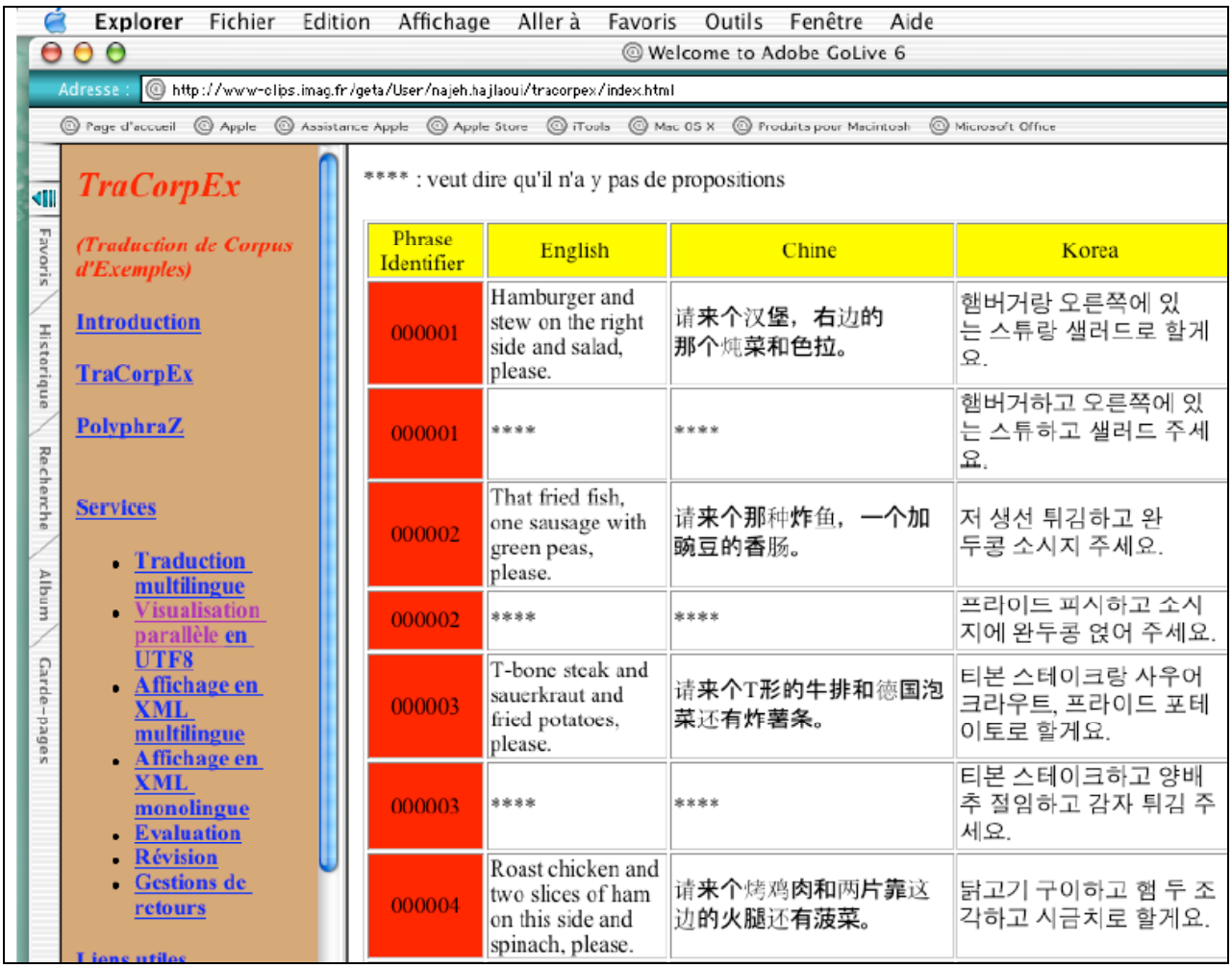

Figure 5: parallel visualisation of the BTEC (extract)

2.6 Evaluation of translation results

We have programmed and integrad in PolyphraZ three evaluation methods (NIST, BLEU and distance calculation). NIST and BLEU are well known. Let us give more details about distance calculation between 2 sentences.

The distance we compute between two strings is a linear combination of two edit distances, one at the level of characters, the other at the level of words. In general, the edit distance between two strings P1 and P2 of atoms (characters or words here) is the minimal number of suppressions, insertions or replacements of atoms necessary to transform P1 into P2 or, equivalently, P2 into P1. To compute the edit distance between P1 and P2 at the level of words, one segments them into words, computes the character distances between words of P1 and words of P2, and then computes the word distance using words as "large characters".

We use the well-known dynamic programming algorithm of (Wagner, Fischer, 1974). To combine the two levels (characters and words), we use the formula:

$D = (\alpha D_{char} + \beta D_{word})/(\alpha+\beta)$; $\alpha + \beta = 1$

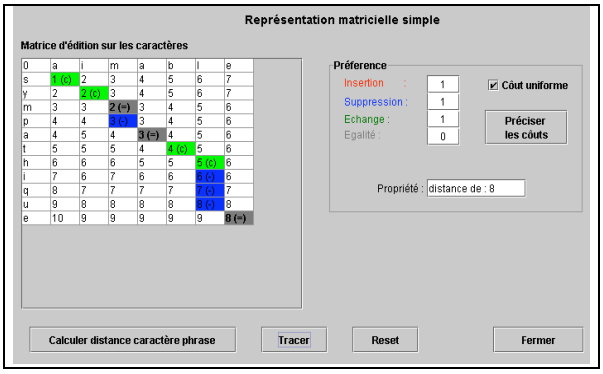

Figure 6:trace of the Wagner and Fischer algorithm

2.6.1 "Track changes" visualisation

This representation corresponds to the presentation used by Microsoft Word in "Track changes" mode. It is very readable. In certain cases, the representation at the level of the characters is more compact and readable that at the level of words, while it is the opposite in other cases. In fact, this

representation is not "faithful" to the trace, because a sequence of exchanges is transformed into a sequence of suppressions and a sequence of insertions.

~~ai~~sympa~~bl~~thique

Figure 7:"Track changes" display

One interesting and today unsolved problem is how to merge the 2 levels: given 2 sentences and their character and word edit distances, necessarily both minimal, how to produce a trace which would be "the best" or "a best" combination of the 2 traces?

2.6.2 Representation with 3 lines

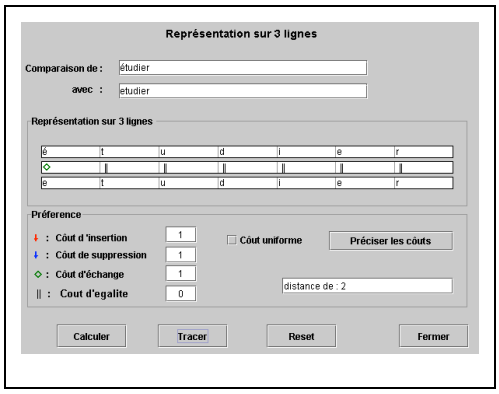

Figure 8 : 3 lines representation

This representation is simpler to understand, but takes more space.

◇ represents the exchange of a character by another,

|| represents the equality between two characters

▯ represent the suppression of the 1st character,

```
<?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 9 : XML representation

3 Conclusion

The CXM and CPXM levels of PolyphraZ are already used. They have allow us to import the BTEC multilingual corpus of parallel sentences (into the common CPM format), to transform it (163000 sentenes in 5 languages) into files in CPXM formats, and to visualize it[1] on the web. The Tanaka corpus should be available when this paper will be presented. The "inner" level of MPM (Multilingual Polyphrase Memory) is almost completed. It will also support versioning.

In the future, we plan to use MPMs not only to handle multilingual corpora of parallel sentences, but also like "pivots", to establish the sentence-level correspondence between parallel monolingual structured documents. If no high quality TWS (like Trados, TM2, Déjà Vu; Transit, etc.) is available, PolyphraZ could be used as a "bare bone" TWS, directly through the web, in the Montaigne[2] spirit.

We are also studying how to integrate into a MPM structure "generators" specifying classes of sentences (automata for messages with variables and variants, regular expressions for CSTAR IF expressions, etc.), and to use them to extend a MPM not only "in width" (addition of new languages), but also "in height", by the automatic creation of new "statements", natural and/or formal.

Références

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

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.

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

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

[1] The full corpus is only accessible to members of CSTAR-III, so that we show only extracts corresponding to parts which are or will be published for the open evaluation of various MT systems to be presented at IWSLT-04.

[2] Mutualization Of Nomadic Translation Aids for Groups on the NEt (Mutualisation d'Outils Nomades de Traduction avec Aides Informatiques pour des Groupes sur le NEt).

H.Vo-trung (Vo-trung, 2004) *Réutilisation de traducteurs gratuits pour développer des systèmes multilingues*, accepté à la conférence RECITAL 2004, avril 2004, Fès, Maroc.

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.

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.

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

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

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.

(C-STAR-III) *C-STAR project*, http://www.c-star.org/

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

(TraCorpEx) projet TraCorpEx

http://www-clips.imag.fr/geta/User/najeh.hajlaoui/tracorpex/index.html

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

Appendices

```
<!-- CXM.dtd  (Common eXample Markup ) is a
  DTD  which  describes the corpora
  (multilingual or monolingual), it is the
  simplest format for imported data.

$Author:  Najeh Hajlaoui
  najeh.hajlaoui@imag.fr
$Date: 2003/12/10 01:28:30 $ -->
<!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>
<!ATTLIST sentence xml:lang CDATA #REQUIRED>

<!ELEMENT sentence (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 and also
  by an attribute which indicates the
  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 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  (Common Parallel eXample
  Markup ) is a  DTD  which  describes the
  multilingual documents (m languages),
  multiversions (n versions) (n>m), it
  allows the description of a collection of
  polyphrases in a single format and
  encoding.
$Author:  Najeh Hajlaoui
  najeh.hajlaoui@imag.fr
$Date: 2003/06/10 01:28:30 $ -->
<!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>

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

<!ELEMENT  monolingual-component
  (segment*) >
<!ATTLIST monolingual-component xml:lang
  CDATA #REQUIRED>
<!ELEMENT  segment (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
  segment in question; it is identified by
  an attribute which indicates the language
  -->
<!-- Segment represents the level of
  alignment, it is usually a sentence -->
```

Appendix 2 : CPXM.dtd (Common Parallel eXample Markup)

```
<!-- MPM.dtd  (Multilingual Polyphrases
  Memory ) is a  DTD  which  allows the
  generation of sentences aligned in
  several languages and the management of
  the correspondence 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-corespondance-
  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
  represents the link of corespondance
  between the diférents proposals of
  translation -->
```

Appendix 3 : MPM.dtd (Multilingual Polyphrase Memory)