CLIPS
Communication Langagière et
Interaction Personne-Système
CNRS - INPG - UJF
BP 53 - 38041 Grenoble Cedex 9 - France

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

Najeh HAJLAOUI & Christian BOITET
GETA, CLIPS, IMAG, Grenoble
University Joseph Fourier, CNRS & INPG

Outline

- Situation
- Introductory example
- Problems
- Solution: building the PolyphraZ tool
  - TraCorpEx and PolyphraZ
  - Objectives
  - Platform (architecture, users)
  - Scenarios for using PolyphraZ
- Conclusion

Situation

- The BTEC corpus of C-STAR : 163000 sentences complete in English (8859-1), Japanese (EUC-JP) and Chinese and Korean => 6.1 Mb for each language
- The TANAKA corpus (English, Japanese) for the Papillon project => 18.4 M.b
- UNL Corpus

- Our goal is to produce a French version

Example from the BTEC J-E

# (copyright) ATR Spoken Language Translation Research Labs.	# (copyright) ATR Spoken Language Translation Research Labs.
# file: 00003_jpn.txt	# file: 00003_usa.txt
# coding: euc-japan	# coding: iso-8859-1
# update: Tue May 21 10:19:57 JST 2002	# update: Tue May 21 10:19:20 JST 2002
# source: jpn	# source: jpn
# target: -	# target: usa
002001＼ケチャップとマスタードはつけますか。	002001**Catsup** or mustard?
002002＼両方つけてください。	002002\Both, please.
002003＼トッピングはどうしますか。	002003\What do you want on it?
002004＼ベーコン、トマト、チーズにしてください。	002004\Bacon, tomato, and cheese, please.
002005＼こちらで召し上がりますか。	002005\Are you going to eat it here?
002006＼ええ。	002006\Yes**, I am**.
002007＼他には何か。	002007\Anything else?
002008＼コーラを。	002008\I would like a Coke.
002009＼小、中、大とありますが。	002009\Small, regular, or large?
002010＼中サイズをください。	002010\Regular, please.
…	…
002512＼寒いので毛布を貸してください。	002512\I feel cold.|Please give me a blanket.
002513＼ご搭乗のみなさま、おはようございます。|本日はエムオー航空六便、ニューヨーク行きをご利用いただきまして、誠にありがとうございます。|この便の機長は上田、私はパーサーの赤坂でございます。|この飛行機は、まもなく離陸いたします。|お座席のベルトはしっかりとお締めでしょうか。|また、禁煙のサインが消えますまでは、おタバコはご遠慮ください。	002513\Good morning ladies and gentlemen.|Captain Ueda and his chief purser Akasaka would like to welcome you aboard our MO Air Lines.|This is flight zero zero six to New York.|At this time, would you please make sure that your seat belt is fastened tight and low, and refrain from smoking until the sign goes off.
002514＼私の席じゃありません。	002514\That's not my seat.

Screen view of revision under Excel

Id	JP	EN	Revision	Systran
JE00597	あっ 、 そう だ 。 保険 に 入っ て おか なく ちゃ 。	Oh, I almost forgot. I should take out some insurance.	Ah, j'ai presque oublié. Je devrais contracter une assurance.	Ah, j'ai presque oublié. Je devrais contracter de l'assurance.
JE00598	緊急 です 。	This is an emergency.	C'est une urgence.	C'est une urgence.
JE00599	ちょっと 考え て み ます 。	I'll think about it, thank you.	J'y penserai, merci.	Je penserai cela, merci.
JE00600	安全 性 が 心配 な 日本人 に 向い た オリエンタル という いい ホテル が あり ます よ 。 コロンビア大学 の 近く です 。	Well, there is a good hotel for security conscious Japanese people called the Oriental. It's located close to Columbia University.	Bien, il y a un bon hôtel pour les Japonais conscients de leur sécurité appelé l'Oriental. Il est situé près de l'université Colombia.	Bien, il y a un bon hôtel pour les japonais conscients de sécurité appelés l'Oriental. Elle est située près de l'université de Colombie.
JE00601	現地 時間 午前 九 時 の 予定 で ござい ます 。	We arrive there at nine a.m. local time.	Nous arrivons là à neuf heures du matin.	Nous arrivons là au temps de gens du pays de neuf heures du matin.
JE00602	日本 の 普通 の サラリーマン や オー エル に 、 そんな 流行 商品 を 好き な とき に 買える 余裕 が ある と 思わ れ ます か 。	In your opinion, can the average Japanese company employee afford buy such trendy goods as they want?	À votre avis, l'employé japonais moyen de compagnie peut-il se permettre l'achat de marchandises aussi dernier cri qu'ils le veulent ?	À votre avis, l'employé japonais moyen de compagnie peut-il se permettre l'achat de telles marchandises dernier cri qu'elles veulent ?
JE00603	ミニ バー を 使い まし た 。 コーク を 二 本 飲み まし た 。	I used the mini bar. I had two cokes.	J'ai employé la mini barre. J'ai eu deux cokéfie.	J'ai employé la mini barre. J'ai eu deux cokéfie.

Time to produce paraphrases by direct translation and MT revision

Corpus	Num_fich	Num_deb	Num_fin	Parag	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

Comparison of revision times in different settings

Input	Tool	Small file (<1000 sentences)	Large file (20000 sentences)
MT	TextEdit	13 mn/p.	N/A
MT	Excel	11□mn/p.	15□mn/p.
Draft human translation	Word	20□mn/p.	N/A

Problems

- Large corpora with different formats, codings
- No available effective tools to manage them
- Tools such as Excel, TextEdit, BBEdit ?
  - Their usage is difficult, limited (64000 lines for Excel)
  - They do not allow sharing, nor editing and visualizing such corpora on the Web
  - To translate corpora, we must break on several parts and send him to a translation server

Solution: building the PolyphraZ tool

- Enlarge corpora horizontally by adding languages and vertically by adding sentences (utterances)
- Prepare synchronized bilingual (+UNL) working versions
  - to give to (voluntary) contributors
- Take an existing ML site as a starting point:
  - reuse the PAPILLON macrostructure (ML lexical data base)
- Use a central "hub" of "polyphrases" to link sentences
  - & handle versioning at that central place
- Prepare a collaborative web environment
  - to give to the (voluntary) contributors
- Always: build tools for handling *large size* corpora

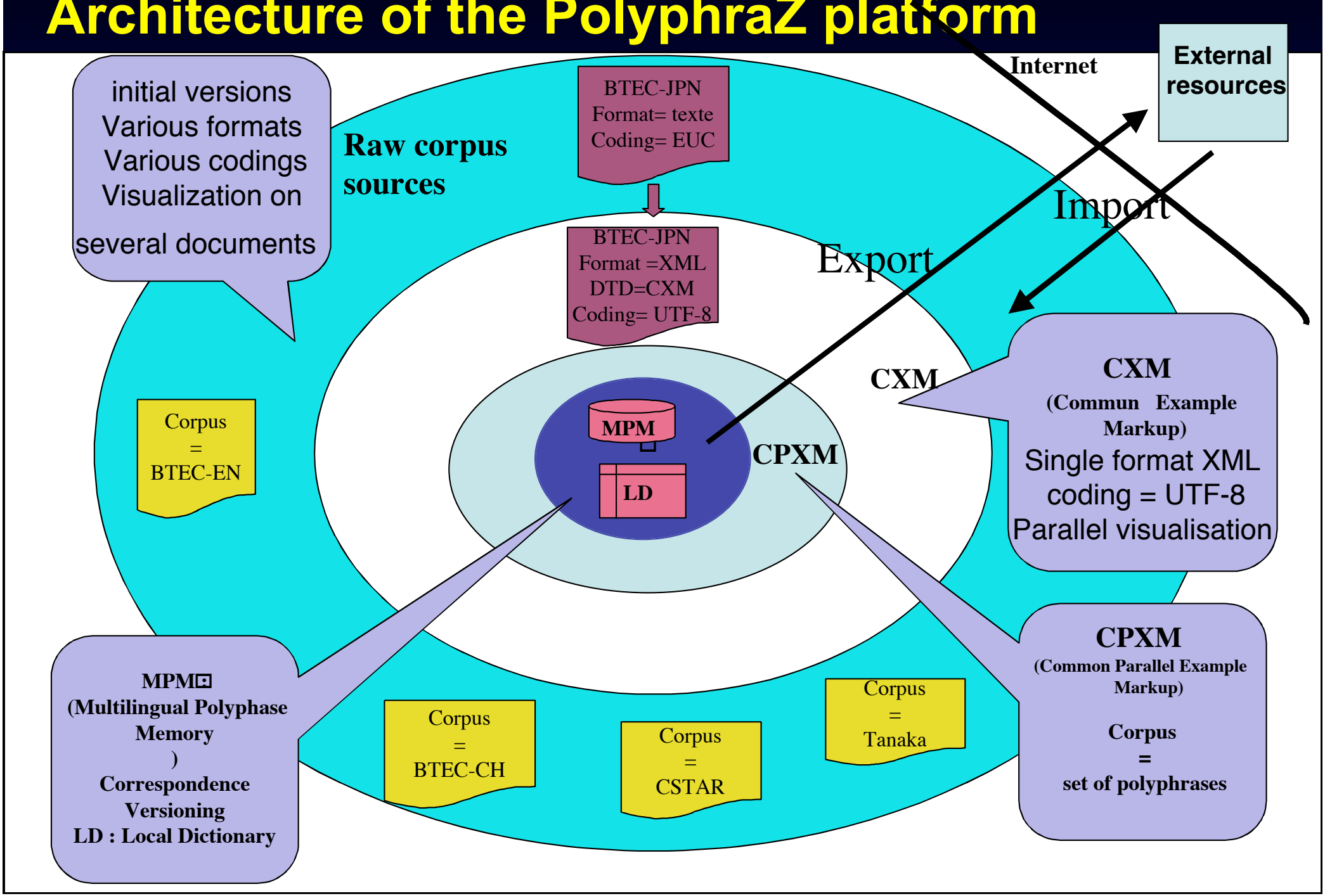
TraCorpEx project and PolyphraZ tool

- The PolyphraZ tool is developed in the framework of the TraCorpEx project (Translation of Corpora of Examples)
- Management of parallel corpora
  - Local
  - On the web
- Monolingual or/and multilingual corpora converted in three format (CXM, CPXM and MPM)

Objectives of TraCorpEx Project

- **Construction of a software platform**
  - Import and export of parallel corpora
  - Preparation of the data and postedition
- **Addition of new languages**
  - French and Arabic for the BTEC corpus and other languages of the Papillon project
- **Evaluation of MT systems**
  - Automatic methods (NIST, BLEU, calculations of distance between sentences and reverse translations)
- **Feedbacks to developers of MT systems**
  - Unknown words, badly translated sentences…

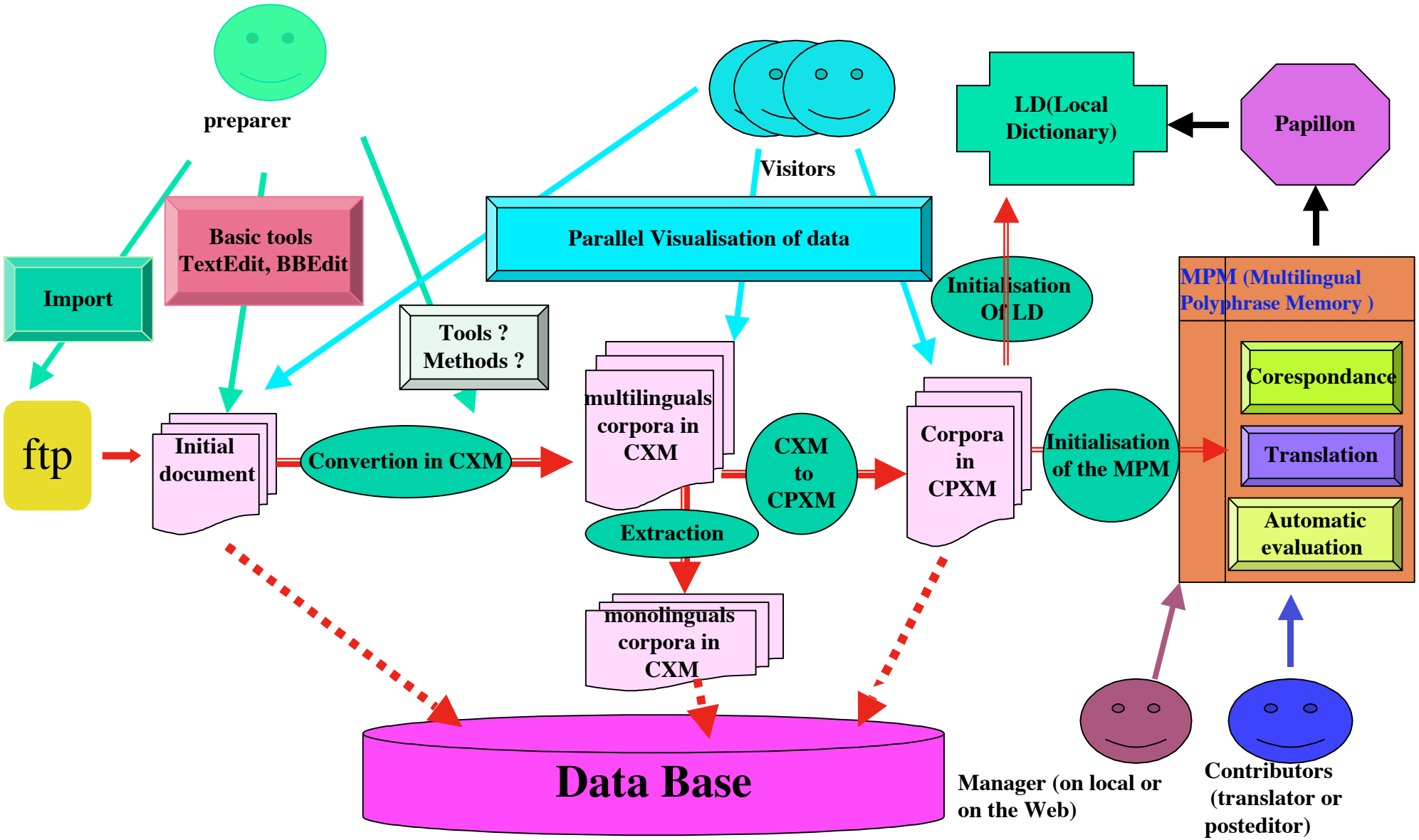
Architecture of the PolyphraZ platform



Users of PolyphraZ

- **The preparer**
  - Calling and parametrizing translation systems, calling evaluation methods and posting the results
- **The reader (normal visitor)**
  - Visualizes the data, various translations, and distances between the character strings
- **The translator-posteditor = contributor**
  - Translates and revises in CPXM and MPM
- **The manager**
  - Produces feedbacks from MPM
  - Proposes suggestions of translation for unknown words
  - Provides a presentation of the evaluations and comparisons between results of various MT systems

Scenarios for using PolyphraZ



Example XML file conforming to the CXM.dtd

```
<?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>
```

**** : veut dire qu'il n'a y pas de propositions

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.

Evaluation of translation results

- **We programmed :**
  - **NIST**
  - **BLEU**
  - **Four presentations based on a distance computation, for subjective evaluation**

Trace of the Wagner and Fisher algorithm

Représentation matricielle simple

Matrice d'édition sur les caractères

0	a	i	m	a	b	l	e
s	1 (c)	2	3	4	5	6	7
y	2	2 (c)	3	4	5	6	7
m	3	3	2 (=)	3	4	5	6
p	4	4	3 (-)	3	4	5	6
a	4	5	4	3 (=)	4	5	6
t	5	5	5	4	4 (c)	5	6
h	6	6	6	5	5	5 (c)	6
i	7	6	7	6	6	6 (-)	6
q	8	7	7	7	7	7 (-)	7
u	9	8	8	8	8	8 (-)	8
e	10	9	9	9	9	9	8 (=)

Préference

Insertion : 1

Suppression : 1

Echange : 1

Egalité : 0

☑ Côut uniforme

Préciser les côuts

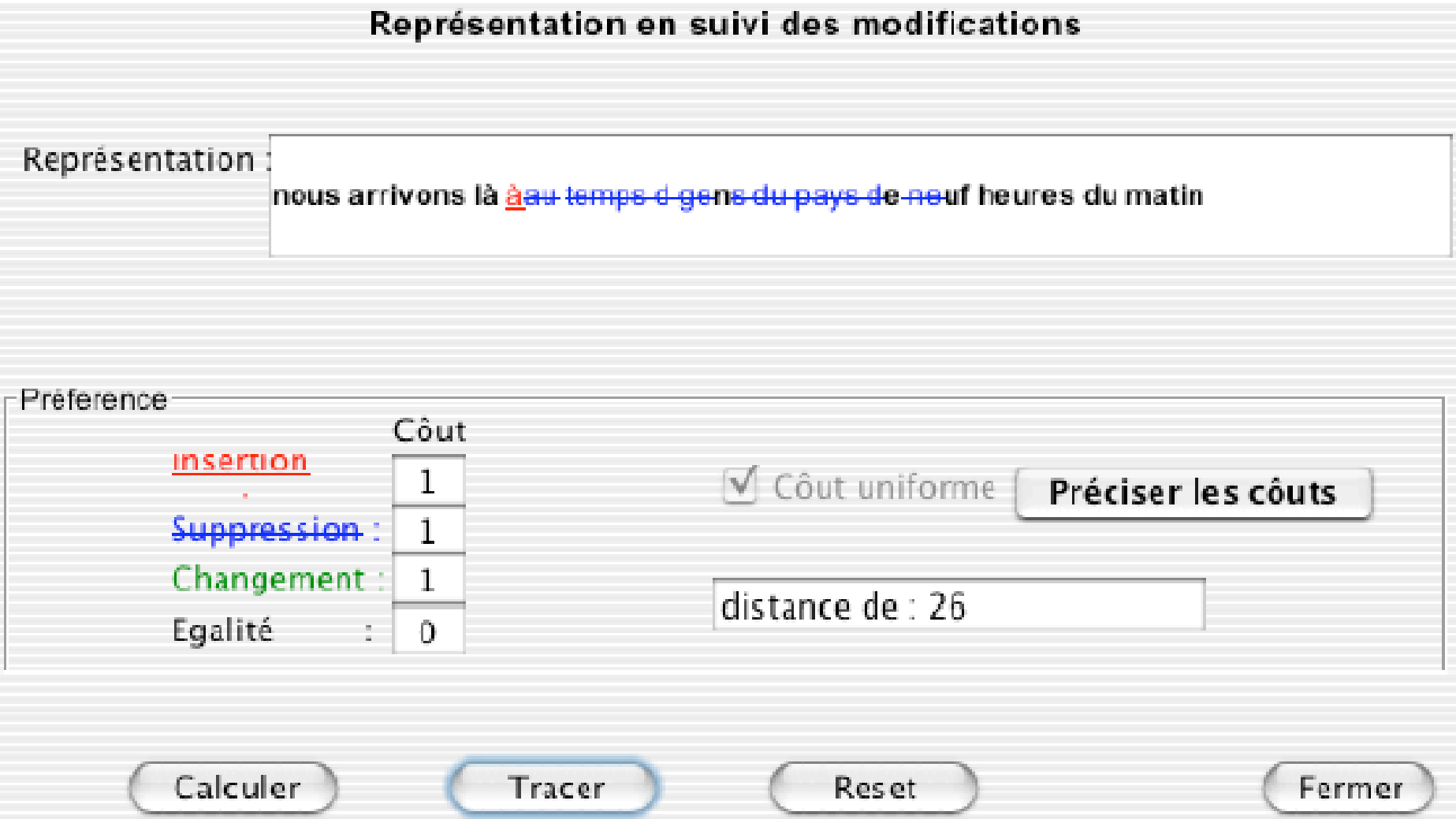
Propriété : distance de : 8

Calculer distance caractère phrase | Tracer | Reset | Fermer

“Track changes” visualisation

Nous arrivons là au temps de gens du pays de neuf heures du matin

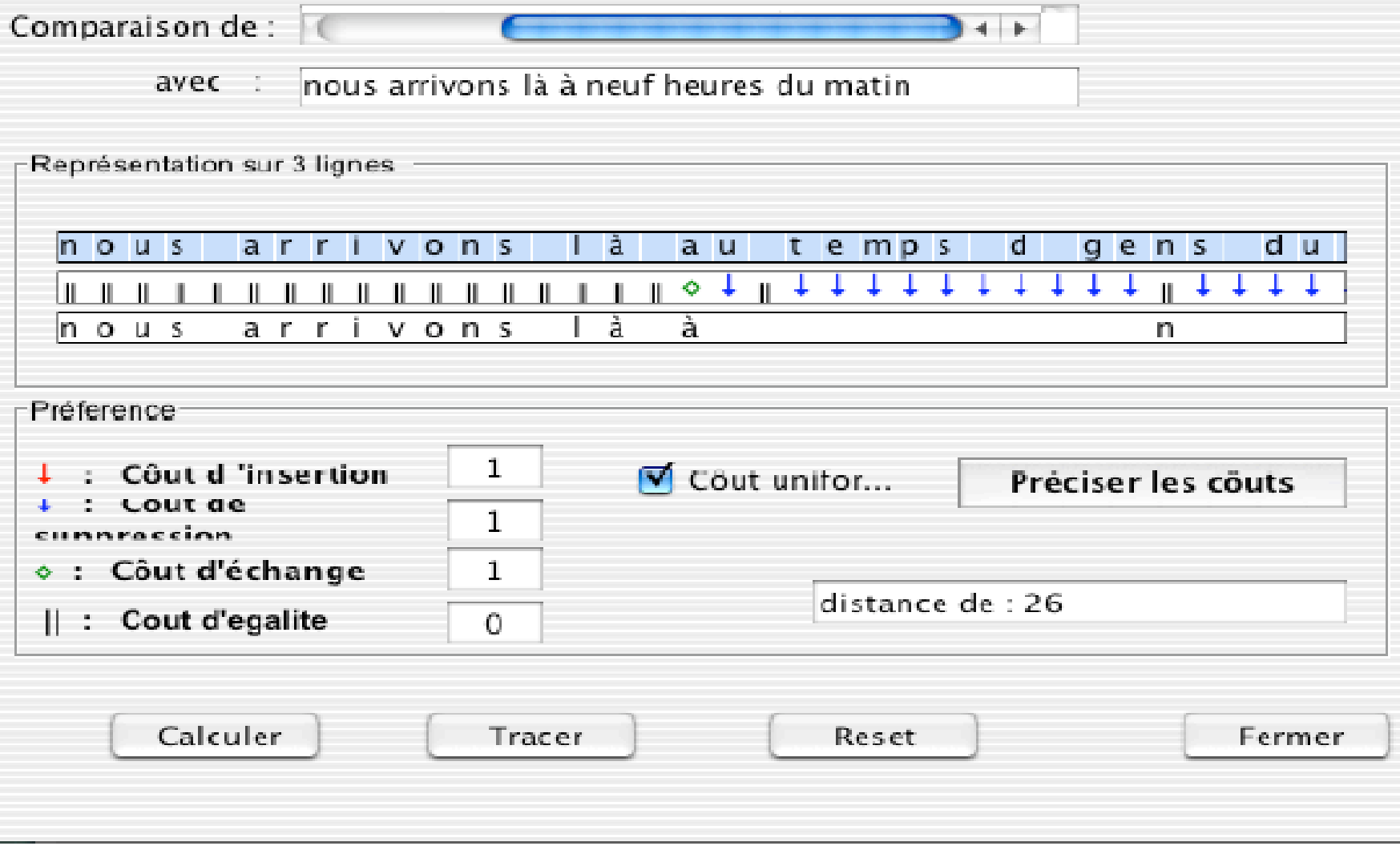
Nous arrivons là à neuf heures du matin.



Representation with 3 lines

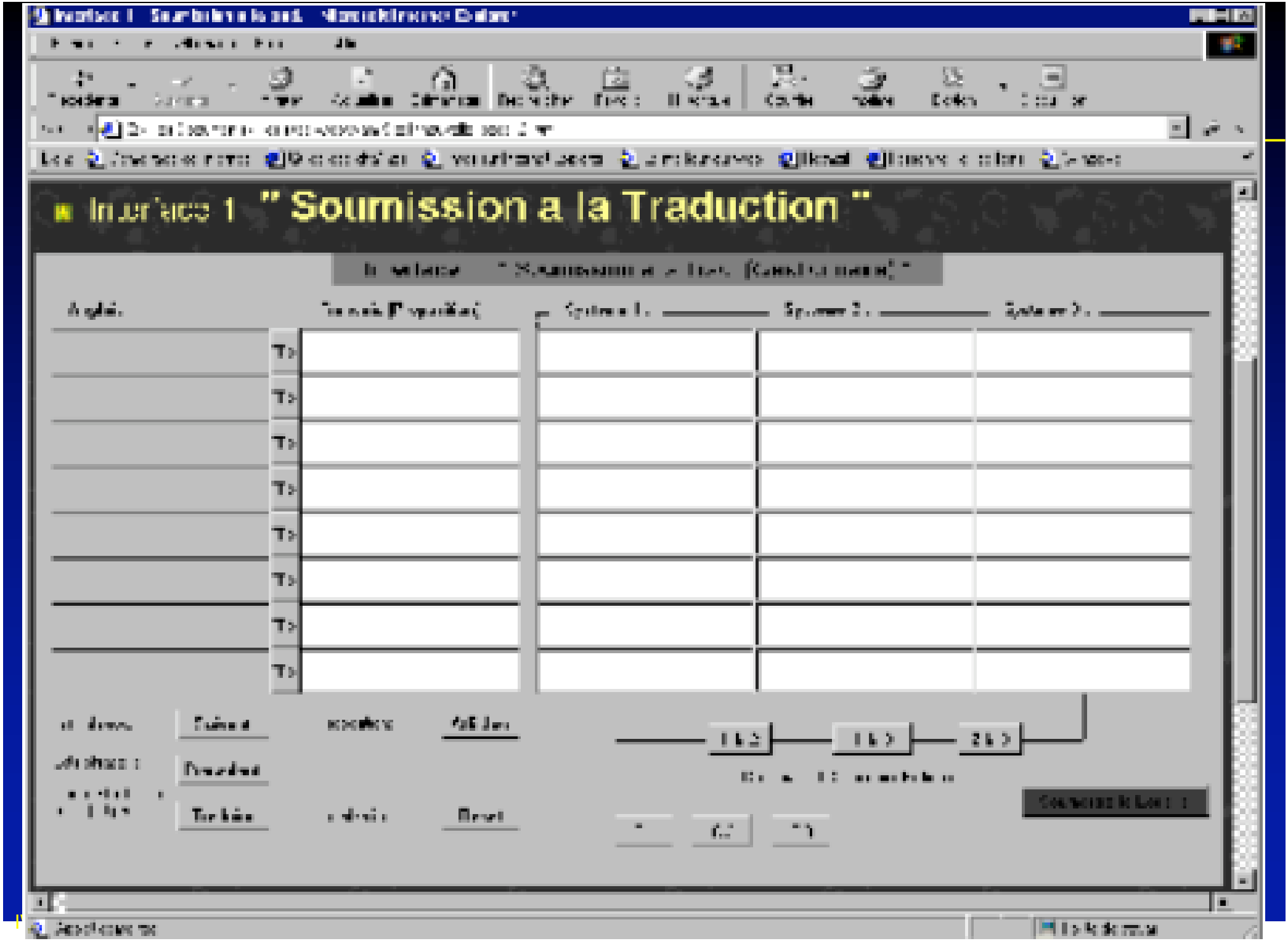
Nous arrivons là au temps de gens du pays de neuf heures du matin

Nous arrivons là à neuf heures du matin.

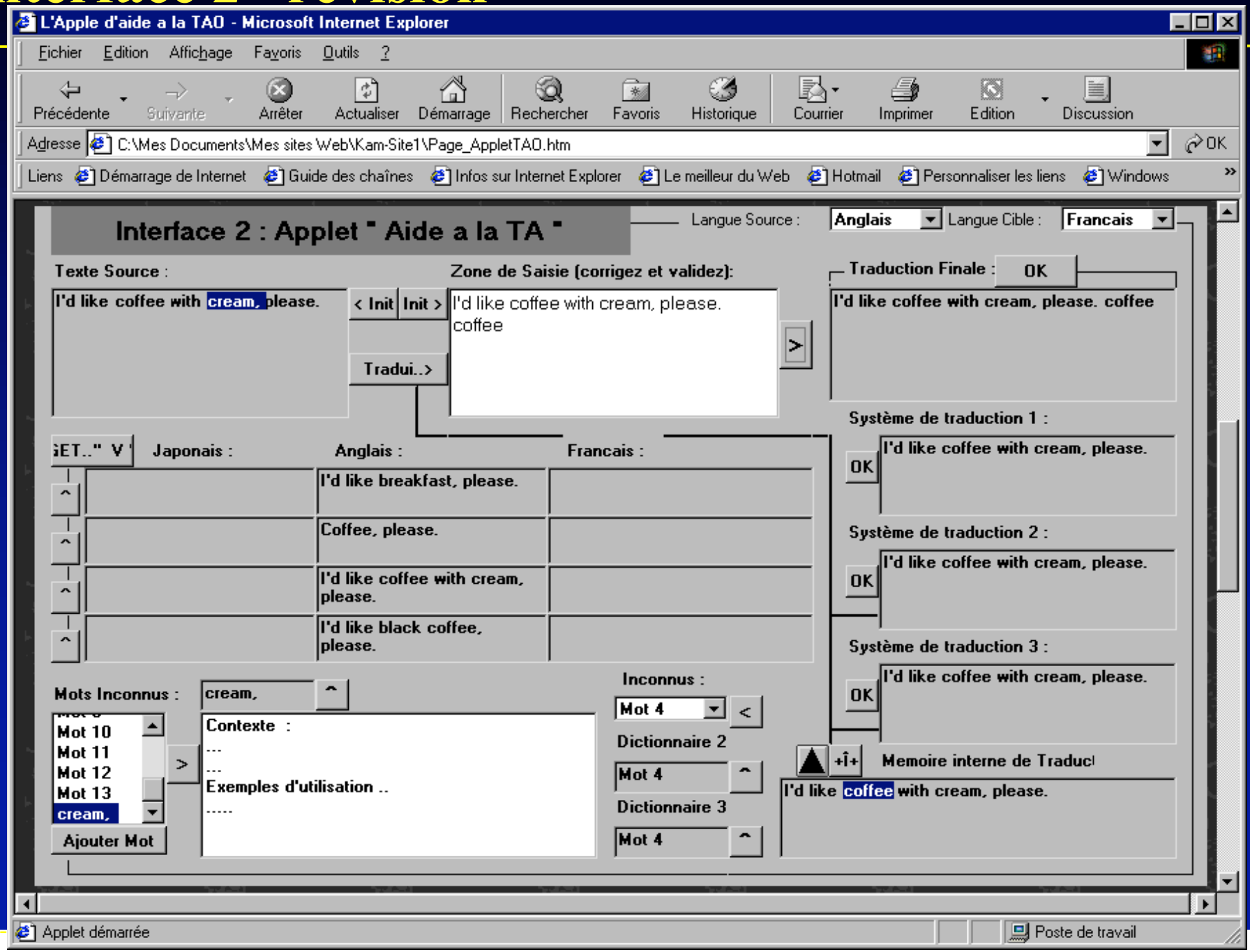


XML representation

```
<?xml version="1.0" ?>
<!DOCTYPE WagnerFischer (View Source for full
doctype...)>
- <WagnerFischer>
  - <Mot>
      <suppression>a</suppression>
      <insertion>a</insertion>
    </Mot>
  - <Mot>
      <insertion>e</insertion>
    </Mot>
  - <Mot>
      <suppression>garco</suppression>
      <insertion>fille</insertion>
      <suppression>n</suppression>
    </Mot>
</WagnerFischer>
```


Interface 1 " Soumission a la Traduction "

Interface 2 “revision”



Conclusion

- **CXM and CPXM levels of PolyphraZ already used**
  - to import the BTEC corpus (in 5 languages) in CXM
  - to transform it into 5-lingual files in CPXM format,
  - to visualize it on the web.

    - Full corpus only accessible to members of CSTAR-III.
- **Plans for the next future:**
  - complete Polyphraz, use it for UNL, Papillon, CSTAR.
  - develop the PolyphraZ web editor
  - use MPMs like « pivots» to establish the sentence level correspondence between parallel monolingual documents.