Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction

Christian Boitet[1], Pushpak Bhattacharyya[2], Etienne Blanc[1], Sanjay Meena[2], Sangharsh Boudhh[2], Georges Fafiotte[1], Achille Falaise[1], Vishal Vacchani[2]

[2] Centre For Indian Language Technology (CFILT)
Indian Institute of Technology (IIT Bombay), Mumbai- 400 076, India
e-mail : {pb,sanjay,sangharsh,vishalv}@cse.iitb.ac.in

[1] GETALP (Study Group for Translation and Processing of Languages and Speech)
LIG (Grenoble Informatics Laboratory), UJF (Univ. Grenoble 1) & UPMF (Univ. Grenoble 2)
e-mail : {Name.Surname}@imag.fr

Abstract

SurviTra is a web service, initially being developed to help a French visitor needing to communicate with an Indian helper when English is not an option. We report on the resource implementation process, which is not trivial, since the multilingual phrasebook contains UNL graphs and natural language sentences possibly containing variables. The dictionary also contains UWs (UNL), which must be found in existing resources, or created, while fitting into the normalization processes of the U++C consortium.

Introduction

Our two laboratories have been cooperating loosely in the framework of the UNL project of multilingual communication (for 12 languages) since 1996, and more actively since 2004, because of the VTHFraDial and CIFLI projects funded by AUF[1] and French Ministries. In the UNL project, each group Gi handles its language, Li, and prepares a UNL[2]-Li "deconverter" and a Li-UNL "enconverter". We want to cooperatively build a common set of UWs, the lexical symbols of UNL, and to agree on ways to represent ("encode") utterances of each Li in UNL. In the VTHFraDial project (2004-05), we collected realistic bilingual task-oriented dialogues, translated by human volunteer interpreters, between French and Vietnamese, Tamil, and Hindi, using our ERIM network-based platform for net-based volunteer interpretation, enhanced with recording functions. We are working on the CIFLI (Indian Languages & French i-Communication) project (2006-08) on building resources and tools for network-based "linguistic survival" communication between French and Indian languages.

The first part of the CIFLI project concerns spoken communication: additional bilingual dialogues are collected, to accumulate data for work on speech translation. The second part concerns written multilingual communication and machine translation (MT) through UNL. To have a concrete applicative goal, we are developing *SurviTra* (Survival Translator), a bilingual chat web service equipped with a phrasebook and a dictionary, both extensible online by users, and enhanced with template sentences and through calls to available MT systems if an utterance is not found using the phrasebook. As UNL-based multilingual translation is our main common research goal, the database behind the phrasebook contains UNL graphs, and the dictionary contains UWs (UNL lexical symbols) associated with specific words (terminology) and fragments of UNL graphs for sentence fragments (phraseology).

The first set of scenarios concern a French visitor in India faced in a "survival situation" (in a taxi, in a restaurant, at the police) and having to communicate with an Indian person, without a satisfactory command of a common language. (Not more than 15% of the people in either country really master English.) Hindi, Marathi, and Tamil are targeted.

In section 1, we outline the functional and the research goals of SurviTra, as well as the specification of its architecture (currently prototyped). In section 2, we describe the methodology for collecting the "corpus", i.e. the phrasebook in this case, and the current results. We do the same for the dictionary in section 3. Some interesting research issues emerge. We conclude by sketching the future work. A demo of the prototype should be possible at the time of the conference.

[1] Association of Universities using French.

[2] Universal Networking Language, www.undl.org.

1 Objectives of building SurviTra

1.1 Functional goals

Situations & constraints. A first decision is whether such a tool should run (1) on the "survivor's" PC or PDA, or (2) on the web, to be accessible from any PC connected to Internet in India. We opt for (2), which is lighter, and has the advantage that the Indian "helpers" will be able to type in their scripts on familiar keyboards (QWERTY, not AZERTY).

A second decision concerns the static or dynamic character of the underlying phrasebook and dictionary. We opt for a dynamic approach. While it is true that paper phrasebook are fixed, as is the spoken phrasebook TalkMan™ (Sony), we think that dynamicity is desirable not only because it is possible, but also because it is necessary if we want to scale up in size and number of languages: users should contribute to the correction and expansion of the data while using the service.

Hence, users should be allowed and even encouraged to correct target as well as source sentences and terms, as well as their UNL counterparts (UNL graphs and UWs), if any.

While using SurviTra, it should also be possible to view and modify data concerning several or all languages supported; hence the need for a relatively "plastic" interface.

Finally, although the first version addresses the problem of communication between French and Indian languages, the service should show English as a supplementary aid, if English is available.

At a first level, the SurviTra window should then have 3 horizontal panes: a chat pane on top, a control pane at the bottom, and a search and contribution pane in between (Figure 1).

Chat pane. The chat function should appear familiar to users. In most chat programs, each user has his own chat window on his PC, and successive turns are marked as to originator and time. Here, there is only one PC, but the same "personalization" should be preserved. Hence, in the chat part of the SurviTra screen, there is one column for the "survivor" (with everything in French); one for the "helper" (with everything in his language); one for the hypothetical "third person" (with everything in English); and one for "the machine", if MT is available. The machine column need show only limited information. For example, if UNL is available, an icon will suffice: the full symbolic text would only be confusing to most users.

The columns should be time-aligned, meaning that the vertical position of the beginning of an utterance is to be interpreted as the time when it was received (from a user, or from the search in the phrasebook, or from an MT system).

Corrections by the users should be carried out simply by editing the text in these columns, using a very simple action (like a right-click) to "unlock" the text of a turn and choose how to modify it. However, there is more to this facility than meets the eye. In particular, a UNL graph should be modifiable indirectly, by "co-editing" it from the text in any language (Boitet & Tsai 2002).

To directly modify a UNL graph, a web-based graph editor should open, preferably at the location of the sentences in the search and contribution pane. Direct graphical modification is actually quite feasible after one or two hours of instruction for a person knowing English at a medium level. If less English is known, the graph may be "localized" to the users language by attaching words to the UW nodes. At any point, to control the graph, the user may invoke UNL deconversions into all languages visible on the screen (where deconversion is available). The results will appear in the chat windows, arranged chronologically.

Search and contribution pane. The left part of the interface (under the chat columns) should show the "found parallel sentences" in the same order as in the chat pane. In templates, the variable names and values can both appear (`$drink=coffee`), or only one may be shown (`$drink` or `coffee`), depending on whether an instance has been found, must be created, or is sent to chat.

The right part of the GUI should contain the dictionary, which should operate in *proactive* mode: rather than ask users to type (or copy and paste) words in a search area, the program should automatically watch the chat columns, so as to segment words, lemmatize them, and look them up in the lexical database, and then show them in some consolidated and useful way in the interface.

In both parts of the GUI, direct manipulation (correction, addition, and perhaps rating or comment) is also necessary.

Domains and control pane. All phrasebooks are organized by domains or situations, sometimes at two levels. Here, we exploit the dynamic character of electronic information to allow users to refine the initial organization, and/or define a new one. The left side of the GUI pane should interface with the domain hierarchy in a graphical way. The right side contains controls (presentation parameters, identification, timing, MT parameters, etc.).

1.2 Research goals

SurviTra is also designed to support and demonstrate research on various approaches to MT. It should actually be an advantage that only a small proportion of the sentences will be "unrecognized" and have to undergo MT, as research-grade MT systems are not usually robust enough to support demanding applications.

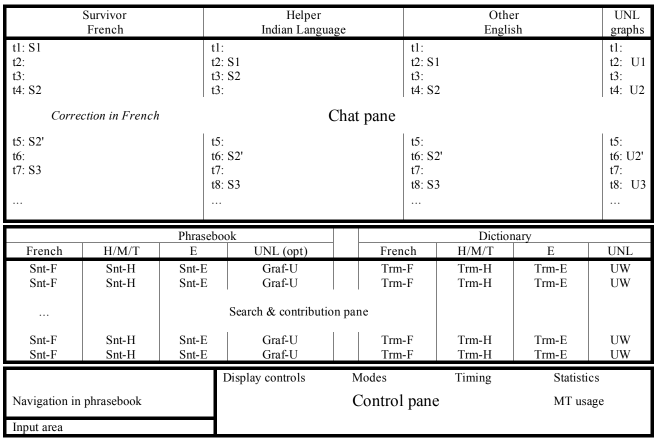
Foremost here is the UNL approach, which necessitates the construction of a common UNL lexical base, deconverters, and enconverters. The lexical database is required before deconversion can be implemented, since deconversion will sometimes need to translate templates containing variables such as nouns, which must be generated in target languages by accessing dictionaries (mostly to retrieve their associated "numerical specifiers"). Deconversion is also useful to check the correctness of UNL graphs stored in the phrasebook.

However, any kind of MT architecture can be used to translate "out of phrasebook" sentences. In particular, the rule-based MT systems currently built by a large national MT project in India under the guidance of Prof. Sinha might be tried experimentally in Indian-Indian survival communication situations. Statistical MT and Example-Based MT systems could also be developed based on the data gathered by the use of SurviTra itself and tried out in parallel with the preceding approaches.

Last but not least, this text-oriented web service could evolve to become multimodal, thus including speech MT.

1.3 Architecture

Interface & scenario



Survivor French	Helper Indian Language	Other English	UNL graphs
t1: S1	t1:	t1:	t1:
t2:	t2: S1	t2: S1	t2: U1
t3:	t3: S2	t3:	t3:
t4: S2	t3:	t4: S2	t4: U2
Correction in French	Chat pane		
t5: S2'	t5:	t5:	t5:
t6:	t6: S2'	t6: S2'	t6: U2'
t7: S3	t7:	t7:	t7:
	t8: S3	t8: S3	t8: U3
…	…	…	…

Phrasebook					Dictionary			
French	H/M/T	E	UNL (opt)		French	H/M/T	E	UNL
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW
…		Search & contribution pane						
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW
Snt-F	Snt-H	Snt-E	Graf-U		Trm-F	Trm-H	Trm-E	UW

Navigation in phrasebook

Input area

Display controls Modes Timing Statistics

Control pane MT usage

Figure 1: SurviTra interface outline

Resource components. These include a multilingual translation memory, a multilingual lexical database, a domain hierarchy, a repertory of users and profiles, and a database containing information (metadata) on usable MT systems (including enconverters and deconverters) and various statistics.

Software components. To support the first-level functions (search, usage, and addition or correction of fixed sentences), we integrate and adapt three web services: a multilingual chat facility (Koine by A. Falaise); the web-based contributive translation support system BEYTrans (Bey, Kageura, Boitet 2006); and the PIVAX (Nguyen 2007) multilingual database for MT systems using a "lexical pivot", e.g. UNL.

Three more components have to be integrated to handle NL and UNL templates, to call MT systems on "not found" sentences, and to edit UNL graphs.

2 Corpus building

The *language resources* should satisfy the needs of the final application, in which 90% of the communicative needs may be covered by the phrasebook elements, and only 10% by MT of "out of phrasebook" sentences or "out of dictionary" terms. Resources should be used symmetrically, so that the same data can be reused for other scenarios such as Indian-Indian communication. (Not nearly everyone in India speaks Hindi!)

The "phrasebook" to be built is a special kind of parallel corpus, or rather translation memory, since it contains the "abstract" UNL language and has variables in its sentences and corresponding UNL graphs. A phrasebook unit consists of an English sentence and its UNL graph, the Hindi translation, and the French translation. The unit is constructed (and later used) symmetrically, with English being possibly used as a "bridge" between developers and users.

Size and sources

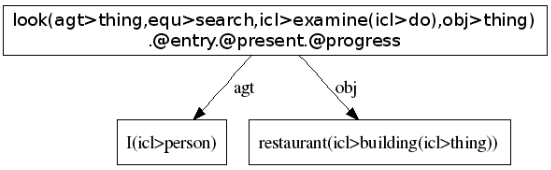
After two months, the phrasebook had about 200 sentences, many with variables ranging over numbers, dates, or word classes (fish, drink, vegetable…). The sentences come from the wikitravel (http://www.wikitravel.org) phrasebooks for Hindi and French, and from the Greenbook[3], intended for French tourists in India.

Example

English: I am looking for a restaurant.

UNL:

[3] Our code name.


look(agt>thing,equ>search,icl>examine(icl>do),obj>thing)
.@entry.@present.@progress
agt
obj
I(icl>person)
restaurant(icl>building(icl>thing))

Hindi: मैं एक रेस्टोरेंट खोज रहा हूँ।

French: Je cherche un restaurant.

UNL graphs

We built about 170 UNL graphs out of 200 sentences. (Graphical images for these graphs are stored separately.) Graphs can have variables, to be replaced by appropriate UWs (e.g. time, place, food item). Graphs were made under the guidance of E. Blanc, author of the French deconverter, so most have been checked and are of good quality.

Desired output

Variety. As for domains, the phrasebook supports: speaking with authority figures, accommodations, money matters, problems (police, theft), common questions, restaurants, shopping, travel, and sentences for acceptance, refusal, or salutations.

We plan to extend the phrasebook with respect to number of languages and domains supported, as well as depth within each domain. Right now, we are focusing on expanding the restaurant domain and adding translations in Marathi and Tamil.

Formats. We used text processors and Excel™ to prepare the corpus. It was then imported in Revolution "stacks" programmed by E. Blanc. For typing Devanagari, we used Baraha7.0 (http://www.baraha.com/) and quillpad (http://quillpad.in/hindi/).

In the future, we will use local tools for preparatory work and a web-based corpus management module (Huynh C. P. PhD work) to collaboratively edit and expand the corpus.

Checking UNL graphs by deconversion

Quality level. The sentences have been verified by native speakers of Hindi and French. We have tried to align sentences where possible, but not to the point of making them sound odd to native speakers.

To test the quality of UNL graphs, a more powerful method than visual inspection is to deconvert them into one or more languages. The French and Hindi deconverters are already working, and give acceptable quality outputs on these UNL graphs. However, differences in UW dictionaries of different language groups still cause problems which are being investigated. (See 3.3 below.)

Methodology, problems, results

Methodology. Our primary emphasis has been on the restaurant domain. A small demo of deconversion and enconversion (at a later stage) is planned concerning this domain.

Variables naturally appear in these sentences. They are now dummies (---) but will soon be typed, as the UW variables above ($hour, $town, $vegi…).

English was used as central language for translation and making of UNL graphs, since English is the pivot language between countries working on this project and U++ as a whole. Moreover, UNL is built upon English, so it makes most sense to convert into UNL from English sentences.

Problems. While translating, we also faced the issue of divergences between the three languages at hand. For example, idiomatic expressions are very often not parallel, and there are differences in politeness levels between English (which has only 'you') and French and Hindi, which are similar in this respect.

Results. The results of this first step of corpus construction are quite encouraging: the desired quality level has been reached, and the quantity, while still about half that of some phrasebooks, enables the presentation of small but convincing demos.

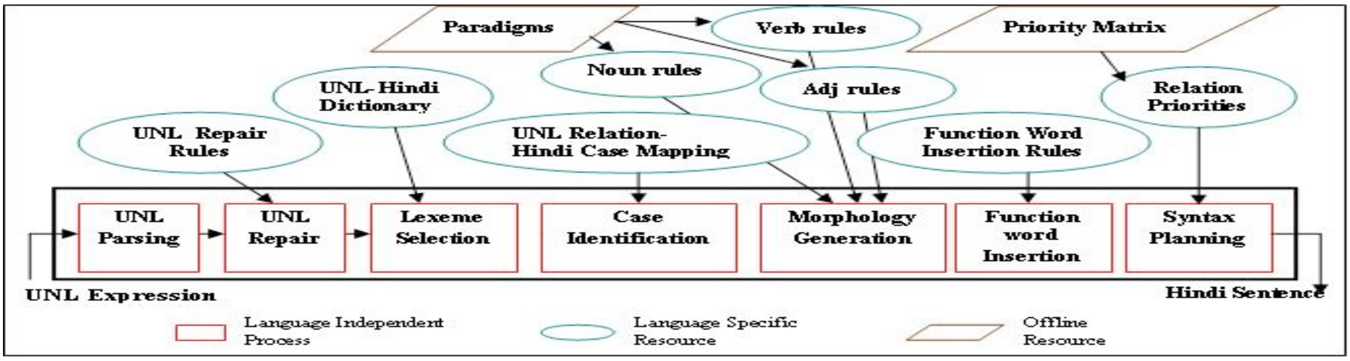


Figure 2 : Hindi Deconversion System

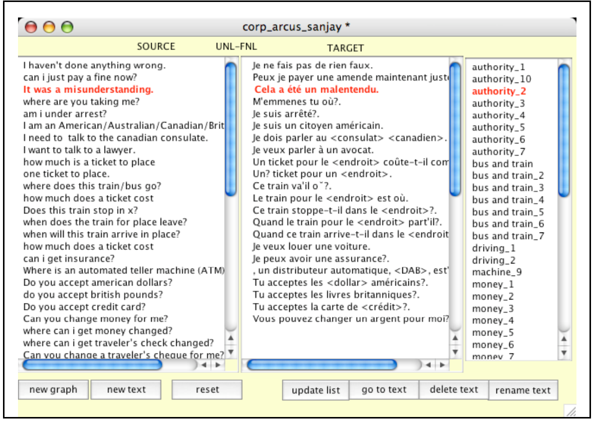
Figure 2 shows the architecture of the CFILT Hindi Deconverter system (HinD). It is rule-based and consists of four main stages: lexeme selection, morphological generation of lexical words, function word insertion, and syntax planning. All of its components use language-independent algorithms operating on language-dependent data. For example, UNL expression parsing and lexeme selection use language-independent algorithms. More details about deconversion, its linguistic resources, and its modules can be found in (Singh & al. 2007).

We have also used E. Blanc's French deconverter to test the UNL graphs produced by the Indian coauthors. As the U++C precise guidelines were followed, the results were quite good, even at this early stage (Figure 3).

3 Dictionary construction, with UWs

3.1 Specifications

The dictionary is implemented so as to be usable for Hindi Deconversion and (later) Enconversion. An entry contains the Hindi lexeme, its corresponding universal word (UW), and its grammatical and semantic features.



I haven't done anything wrong.
can i just pay a fine now?
It was a misunderstanding.
where are you taking me?
am i under arrest?
I am an American/Australian/Canadian/Brit
I need to talk to the canadian consulate.
I want to talk to a lawyer.
how much is a ticket to place
one ticket to place.
where does this train/bus go?
how much does a ticket cost
Does this train stop in x?
when does the train for place leave?
when will this train arrive in place?
how much does a ticket cost
can i get insurance?
Where is an automated teller machine (ATM)
Do you accept american dollars?
do you accept british pounds?
Do you accept credit card?
Can you change money for me?
where can i get money changed?
where can i get traveler's check changed?
Can you change a traveler's cheque for me?

Je ne fais pas de rien faux.
Peux je payer une amende maintenant juste
Cela a été un malentendu.
M'emmenes tu où?.
Je suis arrêté?.
Je suis un citoyen américain.
Je dois parler au <consulat> <canadien>.
Je veux parler à un avocat.
Un ticket pour le <endroit> coûte-t-il com
Un? ticket pour un <endroit>.
Ce train va'il o˘?.
Le train pour le <endroit> est où.
Ce train stoppe-t-il dans le <endroit>?.
Quand le train pour le <endroit> part'il?.
Quand ce train arrive-t-il dans le <endroit
Je veux louer une voiture.
Je peux avoir une assurance?.
, un distributeur automatique, <DAB>, est'
Tu acceptes les <dollar> américains?.
Tu acceptes les livres britanniques?.
Tu acceptes la carte de <crédit>?.
Vous pouvez changer un argent pour moi?

authority_1
authority_10
authority_2
authority_3
authority_4
authority_5
authority_6
authority_7
bus and train
bus and train_2
bus and train_3
bus and train_4
bus and train_5
bus and train_6
bus and train_7
driving_1
driving_2
machine_9
money_1
money_2
money_3
money_4
money_5
money_6
money_7

Figure 3 : stack with French deconversions

An example entry from the dictionary is:

[नाश्ता] {} "breakfast(icl>eat>do, agt>thing)" (N, NOTCH, MALE, INANI, NA);

The first field is the Hindi lexeme displayed in Devanagari font. The second is its Universal word (UW), extracted from the U++C website (http://www.unl.fi.upm.es:8099/unlweb).

The last field contains between parentheses the morpho-syntactic and semantic attributes of the Hindi lemma, which control various generation decisions of the Hindi deconverter, such as the choice of specific case markers.

The capitalized grammatical features belong to three different classes: (a) morphology (gender, number…), (b) syntax (transitive, countable…), and (c) semantics (animate, human…).

3.2 Methodology

Construction of the dictionary. First, we have used existing resources: a Hindi phrasebook available from wiki-travel (http://wikitravel.org/), and the Hindi-French Green book for lexemes pertaining to the *restaurant* domain.

We have then used the UNL-Hindi Deconversion Dictionary (more than 116,000 entries) developed at CFILT (IITB) to obtain the grammatical and semantic attributes of the Hindi lemmas.

Third, the universal words (UWs) were taken from the U++C UW website, which contains about 300,000 UWs built using previous UW sets and enriched by constraints based on WordNet (Miller 1986), e.g. "icl>WN_synonym". Missing Hindi lexemes and UWs were taken from the Hindi-French and UNL-French Lexicons.

UW	English	Hindi	French
toast(icl>bread>thing)	toast	टोस्ट	des toasts

Figure 4: a Hindi-French lexicon entry

The morpho-syntactic and semantic attributes of Hindi words are described in (Singh & al., 2007).

The Universal Words and the grammatical attributes for the lexemes were selected depending upon the domain and the relevance. Proper formatting of the entries in the dictionary was done using a java program. The program takes a text file with Hindi words, their UWs, and grammatical attributes, and produces the dictionary output with proper formatting.

Problems encountered. Complete automation of the dictionary building process was difficult, as selection of UWs and grammatical attributes required human processing.

Many lexemes didn't have any UWs on the U++ website. We used the most relevant ones already listed in the Hindi dictionary.

Results. At the time of writing, we had a dictionary of 350 words pertaining to the Restaurant and Accommodation domains. 1000 are targeted.

3.3 Unification of UW dictionaries with U++C standards

We attempted to unify UW dictionaries using the WordNet hierarchy. For this purpose, we used the three fundamental WordNet proximity notions, i.e *same synset* (closest binding), *same hypernymy hierarchy* (next closest), *sibling* (next).

A few issues remain to be resolved. The restrictions used in UW dictionaries often do not match with the restrictions used in UWs in the U++C database. For nouns, we could search for the closest UW because they had "icl" restrictions. But for verbs and adjectives, search was more difficult, since they usually had no common restrictions which could be used as a parameter for closest match search.

Conclusion

SurviTra is an ongoing project presenting difficult research issues, notably concerning the integration of UNL-based MT to complement the basic phrasebook "linguistic survival" communication service. We have described the initial efforts of data collection, but we expect that data size will grow as users will naturally correct translations and contribute new sentences while using the system.

Acknowledgement

This work has been partly supported by the CIFLI project of the ARCUS program. Thanks to anonymous reviewers for helpful comments, taken into account as far as space allowed, and to Mark Seligman, who kindly improved our English.

References

Bey Y., Kageura K. & Boitet C. (2006) Data Management in QRLex, an Online Aid System for Volunteer Translators. International Journal of Computational Linguistics and Chinese Language Processing, 11/4, pp 349—376.

Blanc E. (1999) PARAX-UNL: a large scale hypertextual multilingual lexical database. Proc. NLPRS '99: the 5th Natural Language Processing Pacific Rim Symposium, Beijing, China, November 5-7, 1999, 4 p.

Blanc É. (2000) From the UNL hyper-graph to GETA's multilevel tree. Proc. MT'2000, Oxford, 18-21 Oct. 2000, British Computer Society, 10 p.

Boitet C. (2002) A rationale for using UNL as an interlingua and more in various domains. Proc. LREC-02 First International Workshop on UNL, other Interlinguas, and their Applications, Las Palmas, 26-31/5/2002, ELRA/ELDA, J. Cardeñosa, ed., pp. 23—26.

Boitet C., Boguslavskij I. & Cardeñosa I. (2007) An Evaluation of UNL Usability for High Quality Multilingualization and Projections for a Future UNL++ Language. In "Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2007)", A. Gelbukh, ed., Springer (LNCS 4394), pp. 361-373. (ISBN-10: 3-540-70938-X Springer, ISSN: 0302-9743)

Boitet C. & Tsai W.-J. (2002) Co-edition of texts and UNL graphs to share text revision across languages and improve MT a posteriori. Proc. ICUKL2002, Goa, 25-29/11/02, 8 p.

Dave S., Parikh J. & Bhattacharyya P. (2001) Interlingua Based English Hindi Machine Translation and Language Divergence. Journal of Machine Translation (JMT), 16/1, pp 251–304. (appeared later: ©2003)

Fafiotte G. & Boitet C. (2003) ERIMM, a platform for assisting and collecting multimedia spontaneous bilingual dialogues. Proc. NLP-KE'03, Beijing, 26-29/10/03, 6 p.

Fafiotte G. (2004) Building and sharing multilingual speech resources, using ERIM generic platforms. Proc. COLING-MLR 2004, Geneva, Switzerland, 28 Aug. 2004, 8 p.

Miller G. A., Beckwith R., Fellbaum C., Gross D. & Miller K. (1986) Introduction to WordNet: An On-line Lexical Database. 86 p. http://wordnet.princeton.edu/5papers.pdf (Revised August 1993)

Nguyen H.-T. & Boitet C. (2007) *Vers un méta-EDL complet, puis un EDL universel pour la TAO.* Proc. TALN-07, Toulouse, 5-8/6/07, ATALA, 10 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 Sept. 1999, Asia Pacific Ass. for MT, J.-I. Tsujii, ed., pp. 220—228.

Singh S., Dalal M., Vachani V., Bhattacharyya P. & Damani O. (2007) Hindi Generation from Interlingua. Proc. Machine Translation Summit (MTS 07), Copenhagen, September, 2007, 8 p.