

# Experiments in Distributed MT

Oliver Streiter, Antje Schmidt-Wigger,  
Ursula Reuther and Catherine Pease<sup>1</sup>

IAI

Martin-Luther-Straße 14  
66111 Saarbrücken Germany  
catler@iai.uni-sb.de

*Connecting two NLP modules which have been developed at different sites required until recently that one system had to be frozen in its current state and then transferred and linked by an interface to the second system. With the use of the Internet it has now become possible to keep both NLP modules at their sites, where they are maintained and updated independently. The total translation process then is distributed across the sites and the Internet is part of the interface. In this paper we report on 3 experiments undertaken wrt distributed Mt and the underlying harmonization and sharing of linguistic resources*

## 1 Introduction

The construction of an MT system requires a huge effort which involves the cooperation of different specialists, such as computer scientists, linguists, translators and lexicographers. Once an MT system has been developed, there may be a need to extend the system by additional target or source languages. At this point the developers have two options, (1) to integrate a new language component into the existing system, which requires the collaboration with new specialist for a period of several years, or (2) to use other existing analysis or generation systems and to connect the two systems. The feasibility of this latter approach has been explored for example in [Schütz89] and [Mesli94], where the CAT2 MT-system has been linked to the Penman system. For this purpose one system has been frozen in its current state and has then been linked to the second system. Further maintenance of the frozen system, however, became impossible.

In the project presented here, "CAT2: Traducción Automática Multilingüal"<sup>2</sup>, in the frame of which two MT systems, located at the University of Mexico City (UNAM) and at the IAI in Saarbrücken building on the same formalism have been linked, thus pursuing the above mentioned second approach, maintaining however the two systems independently at different sites and connecting them via the Internet.

This means that not only two systems are going to be developed and maintained at different sites, but the whole process of translation is distributed

---

<sup>1</sup>Presented at the Workshop on Distributing and Accessing Linguistic Resources, Granada, Spain - May 27 1998

<sup>2</sup>This project has been jointly sponsored by CONACYT (Consejo Nacional des Ciencia y Tecnología in Mexico) and the Forschungszentrum Jülich GmbH in Germany within the frame of a mutual agreement on cooperating in the domain of research and technology.

over two continents. In the remainder of this article, we present the two systems which have been linked via the Internet, the architecture used for translation and the experiments undertaken to match and share linguistic resources.

## 2 The MT systems

CAT2 is an NLP formalism developed for the purpose of multilingual MT (cf. [Sharp94], [Sharp and Streiter95]). Within this formalism different grammars and lexicons have been developed within different projects by different groups of researchers. A group of research institutions from Belgium, Luxembourg and Germany developed the ANTHEM system the purpose of which is the translation of medical diagnostic expressions (cf. [Ceusters et al. 94], [Streiter and Schmidt-Wigger95a]). A consortium consisting of the IAI, the University of PARIS VII and the ERI in Cairo developed an English-German-Arabic MT component for medical classifications (cf. [Pease and Boushaba96]). At the UNAM (Universidad Autónoma de México), a bi-directional English-Spanish MT system has been developed mainly for the purpose of translating newspaper articles (cf. [Sharp and Streiter95]). A German-English-French MT component has been developed at IAI Saarbrücken, which, similar to the UNAM system, is not tuned to special subject domains or sub-languages (cf. [Streiter et al. 94], [Streiter96]) (<http://www.iai.uni-sb.de/cat2/cat/en/trans.html>). All systems can be started from a web-page, where the translation is effected either interactively or in batch mode. In the latter case the translation is delivered to the user via e-mail. The translation service is free. In order to offer new language pairs to the user, UNAM and IAI decided to link their MT systems while keeping them independent for further development at their respective sites.

## 3 Linguistic Interface

In order to link the two MT systems, a linguistic interface had to be written. This interface has to cope on the one hand with differences between languages and on the other hand with differences between the conception of the two MT-systems. It consists of three sub-components, the lexical transfer rules, the feature transfer rules and complex translation rules. While the lexical transfer rule mainly accounts for differences between the languages (e.g. the lexicon), the latter two account mainly for differences between the two MT components. Whenever it was possible, feature transfer and structural transfer were avoided by a mutual harmonization of linguistic structures. By negotiating which of the two translation strategies had to be followed, mutual understanding of the other's MT strategy became possible. For each of these sub-components we shall give one example.

### 3.1 Lexical Transfer Rules

Lexical transfer rules represent the kernel of the translation process. The following is a sample of some transfer rules of the transfer component linking the UNAM Spanish analysis component with the IAI French generation component.

UNAM	IAI
t = {lex=abarcar}	<=> {lex=comprendre}.
t = {lex=abastecer}	<=> {lex=fournir}.
t = {lex=abertura}	<=> {lex=ouvrir,lemma=ouverture}.
t = {lex=abierto}	<=> {lex=ouvrir,lemma=ouvert}.
t = {lex=abrir}	<=> {lex=ouvrir,lemma=ouvrir}.

What should be noted here is the difference in representation of derived words. In the IAI system, a word family is grouped through one lexeme value, usually a verb or noun. The choice of the surface realization is made through extensive semantic encoding in the lexicon (cf. [Streiter and Schmidt-Wigger95b]). In the following entry, for example, the choice of the lemma of the target language is constraint by the semantic values associated with every lemma:

```
IAI

ouvrir=
{lex=ouvrir}&
({lemma=ouvrir,head={AVOIR,ehead={sem={EVENT}}}}
;{lemma=ouvreur,head={CNT,ehead={MASC,sem={NOPROF}}}}
;{lemma=ouvreuse,head={CNT,ehead={FEM,sem={NOPROF}}}}
;{lemma=ouvert,head={ADJ_POST,ehead={sem={STATE}}}}
;{lemma=ouvertement,head={ADV,ehead={sem={EVENT}}}}
;{lemma=ouverture,
  head={CNT,ehead={FEM,sem=({APERTURE};{EVENT})}}})&
{sc={a={AGENT},b={THEME}}}
```

As the semantic encoding in the UNAM lexicons is less fine-grained, the lexical choice cannot be made on this basis only. As a consequence, the additional specification of the lemma itself is necessary in the transfer rules as e.g. in the transfer rules for *abertura* and *abierto* above.

### 3.2 Feature Translation

In order to give an example of feature translation, we shall show the transfer of modality between the two systems. The information content of modality is well described in literature and the differentiation is thus made in an equal way by the UNAM and by the IAI grammar. But besides different names for the mood values and attributes, the feature is integrated differently in

the whole hierarchically organized feature structure. While in the UNAM grammar, mood just follows the principle of ‘extended head’-percolation (as described e.g. in [Grimshaw91] and [Streiter96]), the IAI grammar introduces it in a complex semantic structure, where mood can only appear within temporal structure (`tmp={}`), which itself is embedded into the set of abstract entities (`abs={}`). This difference had to be handled by means of the following feature transfer rules, where the left-hand side matches the UNAM system and the right-hand side the IAI system.

UNAM	IAI
<code>dec = {head={ehead={mode=decl}}}</code> <=> <code style="padding-left: 40px;">{head={ehead={sem={abs={tmp={mood=d}}}}}}.</code>	
<code>int = {head={ehead={mode=interrog}}}</code> <=> <code style="padding-left: 40px;">{head={ehead={sem={abs={tmp={mood=i}}}}}}.</code>	
<code>exc = {head={ehead={mode=exclam}}}</code> <=> <code style="padding-left: 40px;">{head={ehead={sem={abs={tmp={mood=o}}}}}}.</code>	

### 3.3 Structural Translation

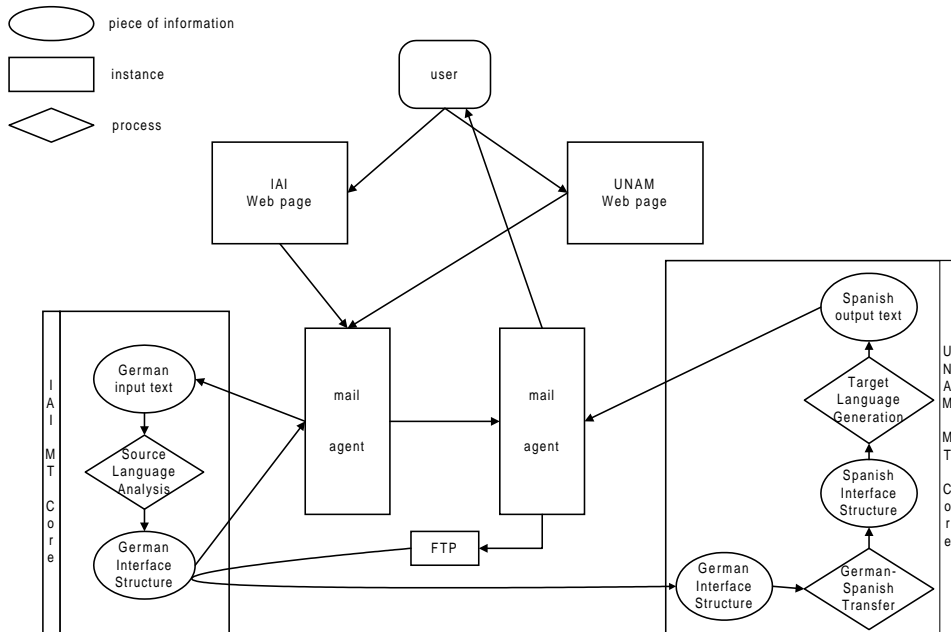
An example of structural translation involves a basic decision about the shape of the interface structure. According to a generally adopted approach to the definition of such an interface structure, the members of a phrase are placed into a canonical order with the head of the phrase in first position. The head and the other members, have then to be re-arranged in generation. The UNAM interface structure follows this principle, while the IAI follows a different strategy. Here, the interface structure reflects word order of the input sentence for allowing better results when analysis fails during transfer, as in such a case the verb of a sentence, for example, would appear in the unnatural sentence-initial position.

An adaptation of the generation principles would result in a complete reshaping of the analysis and generation components of one partner. Thus, through transfer, the structure has to be reshaped to the form awaited for by the generation component. In the following example, the head of the German structure on the right side is placed in the first position on the left side (the Spanish target structure). The markers `a` and `c` represent calls to other transfer rules, so that all constituents before the `head` and after the `head` are transferred and appear in the place indicated on the target side.

UNAM	IAI
<code>change_pos_pred =</code> <code>{}. [head, *a, *c]</code>	<code>&lt;== {role=ROLE, head=HEAD}}. [</code> <code style="padding-left: 40px;">*a,</code> <code style="padding-left: 40px;">head: {min=yes, role=ROLE, head=HEAD},</code> <code style="padding-left: 40px;">*c].</code>

## 4 Architecture of the Translation

Both project partners, UNAM and IAI, added to their web-page for the duration of the test phase the, until then not covered, language pairs Spanish-German, Spanish-French, German-Spanish and French-Spanish. If the SL is Spanish, the input text and the user's e-mail address are sent to the UNAM translation module, if the SL is German or French, the text and the reply address are sent to the IAI translation module.



Every module, however, effects only one part of the whole translation process, i.e. the analysis of the SL, and transforms it into the usual Interface Structure. The result of the analysis and the reply address then are stored as an object file and transferred via ftp to the partner system, when it has been informed via e-mail of the successful analysis of the text (i.e. the German objects are transferred to UNAM and the Spanish objects are transferred to IAI). At the target site the objects are loaded, transferred to the target language and generated. Finally, the generated output is sent to the reply address, as exemplified in the above diagram for the language pair German  $\rightarrow$  Spanish.

This architecture, however, implies that the transfer component which functions as the interface between the two systems and translates from Spanish to German and from German to Spanish has to be available at two different sites, which makes the updating of the transfer component complicated. It would be preferable, to have the transfer component in one copy only at one site (e.g. Mexico) and to receive this transfer component either within a server architecture or attached to the (Spanish) object to be translated at IAI.

## 5 Experiments in Distributed MT

### 5.1 Transfer Approach

Adding new language pairs in the way described led quickly to a stable test version which can handle almost the same amount of structures as can be handled by the two different systems, even if the linguistic modelling is somewhat different. In our first experiment we linked the German and French language modules of IAI to the Spanish language modules of UNAM, using all three transfer components mentioned above, i.e. lexical transfer, feature translation and structure translation.

To get a useful translation service, however, lexical work on transfer rules continues to represent a major part of the work to be done. To circumvent this work, the alternative is to use modules of a language common to both systems, i.e. the English modules, as a pivot. Such a pivot-based approach in distributed MT-environment has already been proposed in [Schubert88].

### 5.2 English as Pivot

Contrary to the proposal of [Schubert88], we do not use an (English) surface string as pivot, but the English interface structure, which is less ambiguous than the related surface string. We thus linked the German and French language modules of IAI to the Spanish language modules of UNAM via the respective English interface structure. Within this approach only feature translation and structure translation has to be accounted for, which are the same as in the previous experiment. Ideally lexical transfer consists of only one line:

UNAM	IAI
$t = \{\text{lex}=\text{LEX}, \text{head}=\{\text{cat}=\text{CAT}\}\}$	$\Leftrightarrow \{\text{lex}=\text{LEX}, \text{head}=\{\text{cat}=\text{CAT}\}\}$

The realization of this strategy resulted in an increased coverage wrt to the previous transfer approach. In addition, it allows for an interesting fall-back strategy in cases where there are mismatches between the two English lexicons or an entry is missing in the target lexicon. In such cases the English source entry can serve as a basis for the creation of a temporary English target entry. Before realizing such an on-line entry conversion, we carried out an experiment off-line to convert the English lexicons in both directions.

### 5.3 Sharing of Lexical Resources

As a preparation for the on-line conversion of English lexical entries, we first attempted to convert the English lexical entries of both institutions into the representation of the partner institution. This was done by providing an interface for feature conversion and then combining the latest, converted

versions of the two lexicons in a database, making them usable for both IAI and UNAM. To give an example of this conversion, the following UNAM lexical entries have been converted into (underspecified) IAI entries using this method:

```
UNAM:      musician=
           {lex=musician,mPERSON}

INTERLEX:  musician=
           {lex=musician,COUNT_NOUN,PERSON}

IAI:      musician=
           {lex=musician,head={COUNT,ehead={sem={HUMAN}}}}

UNAM:      sociology=
           {lex=sociology,mMASS,mCONCEPT}

INTERLEX:  sociology=
           {lex=sociology,MASS_NOUN,CONCEPT}

IAI:      sociology=
           {lex=sociology,head={MASS,ehead={sem={ABSTRACT}}}}

UNAM:      know=
           {lex=know,mVERB,mNONSTATIVE,
            frame={arg1={mAGENT,mSENTIENT},arg2={mTHEME}}}

INTERLEX:  know=
           {lex=know,VERB,NONSTATIVE,
            frame={arg1={AGENT,NOUN,SENTIENT},arg2={THEME}}}

IAI:      know=
           {lex=know,head={VERB,sem={STATE}},
            sc={a={AGENT,head={ehead={sem={HUMAN}}}},b={THEME}}}
```

As shown above, the interface comprises an 'interlexical' representation, which is intended to be a system independent description of the semantic and syntactic information contained in the entries of the different lexicons. Its functioning relies on string substitutions, where each lexicon is connected to an interface which converts entries to and from the interlexical representation.

At present this component is only tailored towards a treatment of the two lexicons involved, but can be relatively easily adapted to other lexical resources.

The next step, not yet realized, would be the on-line conversion during lexical transfer in case where no equivalent can be found in the partner's English lexicon. Such a conversion can be easily integrated into the CAT2 English-to-English translation module, where a special rule type (the so-called r-rules) apply only in case no translation can be found.

## 6 Conclusions

We have shown that the use of the Internet offers new perspectives for the development and use of linguistic resources in general and MT applications specifically. Based on simple Internet facilities such as mail and ftp we installed and tested a translation service which covered new translation pairs (German-Spanish and French-Spanish) where the whole process of translation is distributed over two continents. In two experiments the architecture based on an English disambiguated interface structure proved to be most promising. The deplorable fact that UNAM closed its translation service some months after the end of the project does not invalidate our approach but shows that for such applications not only technical solutions and standardized exchange formats are needed, but more importantly also long-term political and administrative stability in order to convert the new technological possibilities into real-life applications.

## References

- [Ceusters et al. 94] Werner Ceusters, Guy Deville, Emmanuel Herbigniaux, Pierre Mousel, Oliver Streiter, and Geert Thienpont. 1994. The AN-THEM Prototype. IAI WP 31. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Grimshaw91] Jane Grimshaw. 1991. Extended Projection. Brandeis University, Waltham MA 02254, ms, July.
- [Mesli94] Nadia Mesli. 1994. Interlingua vs. transfer? knowledge sharing across projects. In *Technology Partnerships for Crossing the Language Barrier*, Columbia, Maryland, USA, 5-8 October. Proceedings of the First Conference of the Association for Machine Translation in the Americas.
- [Pease and Boushaba96] Catherine Pease and Abd Al-Aziz Boushaba. 1996. ARAMED. Extension and integration of Arabic lingware components in a unification-based MT system for the field of medical terminology and classification. In *First KFUPM Workshop on Information & Computer Science (WICS)*, Dhahran, June 9.
- [Schubert88] Klaus Schubert. 1988. The architecture of DLT - interlingua or double direct? In Dan Maxwell, Klaus Schubert, and Toon Witkam, editors, *New Directions in Machine Translation*. Foris Publication, Dordrecht - Holland.
- [Schütz89] Jörg Schütz. 1989. Towards a Framework for Knowledge-based Machine Translation. IAI WP n.10.

- [Sharp and Streiter95] Randall Sharp and Oliver Streiter. 1995. Applications in Multilingual Machine Translation. In *Proceedings of The Third International Conference and Exhibition on Practical Applications of Prolog, Paris, 4th-7th April*. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Sharp94] Randall Sharp. 1994. CAT2 Reference Manual Version 3.6. IAI WP n.27, IAI, Institut der Gesellschaft zur Förderung der angewandten Informationsforschung e.V. an der Universität des Saarlandes. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter and Schmidt-Wigger95a] Oliver Streiter and Antje Schmidt-Wigger. 1995a. The integration of linguistic and domain specific knowledge: CAT2 within ANTHEM. In *Proceedings of the Conference on Health Telematics95*, pages 387–392, Ischia, July 2-6. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter and Schmidt-Wigger95b] Oliver Streiter and Antje Schmidt-Wigger. 1995b. Patterns of derivation. In *TMI-95*. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter et al. 94] Oliver Streiter, Randall Sharp, Johann Haller, Catherine Pease, and Antje Schmidt-Wigger. 1994. Aspects of a unification based multilingual system for computer-aided translation. In *AVIGNON-94*. URL: <http://www.iai.uni-sb.de/en/cat-docs.html>.
- [Streiter96] Oliver Streiter. 1996. *Linguistic Modeling for Multilingual Machine Translation*. Informatik. Shaker Verlag, Aachen.