# Deep Linguistic Multilingual Translation and Bilingual Dictionaries

**Eric Wehrli, Luka Nerima & Yves Scherrer**
LATL-Department of Linguistics
University of Geneva
{Eric.Wehrli, Luka.Nerima, Yves.Scherrer}@unige.ch

## Abstract

This paper describes the MulTra project, aiming at the development of an efficient multilingual translation technology based on an abstract and generic linguistic model as well as on object-oriented software design. In particular, we will address the issue of the rapid growth both of the transfer modules and of the bilingual databases. For the latter, we will show that a significant part of bilingual lexical databases can be derived automatically through transitivity, with corpus validation.

## 1 Introduction

The goal of the MulTra project is to develop a grammar-based translation model capable of handling not just a couple of languages, but potentially a large number of languages. This is not an original goal, but as 50 years of work and investment have shown, the task is by no means an easy one, and although SMT has shown fast and impressive results towards it (e.g. EuroMatrix), we believe that a (principled) grammar-based approach is worth developing, taking advantage of the remarkable similarities displayed by languages at an abstract level of representation. In the first phase of this project (2007-2009), our work has focused on French, English, German, Italian and Spanish, with preliminary steps towards Greek, Romanian, Russian and Japanese.

To evaluate the quality of the (still under development) system, we decided to join the WMT09 translation evaluation with prototypes for the following language pairs: English to French, French to English and German to English. In this short paper, we will first give a rough description of the MulTra system architecture and then turn to the difficult issue of the bilingual dictionaries.

The MulTra project relies to a large extent on abstract linguistics, inspired from recent work in generative grammar (Chomsky, 1995, Culicover & Jackendoff, 2005, Bresnan, 2001). The grammar formalism developed for this project is both rich enough to express the structural diversity of all the languages taken into account, and abstract enough to capture the generalizations hidden behind obvious surface diversity. At the software level, an object-oriented design has been used, similar in many ways to the one adopted for the multilingual parser (cf. Wehrli, 2007).

The rapid growth of the number of transfer modules has often been viewed as a major flaw of the transfer model when applied to multilingual translation (cf. Arnold, 2000, Kay, 1997). This argument, which relies on the fact that the number of transfer modules and of the corresponding bilingual dictionaries increases as a quadratic function of the number of languages, is considerably weakened if one can show that transfer modules can be made relatively simple and light (cf. section 2), compared to the analysis and generation modules (whose numbers are a linear function of the number of languages). Likewise, section 3 will show how one can drastically reduce the amount of work by deriving bilingual dictionaries by transitivity.

## 2 The architecture of the MulTra system

To a large extent, this system can be viewed as an extension of the Multilingual Fips parsing project. For one thing, the availability of the "deep linguistic" Fips parser for the targeted languages is a crucial element for the MulTra project; second, the MulTra software design matches the one developed for the multilingual parser. In both cases, the goal is to set up a generic system which can be redefined (through type extension and method redefinition) to suit the specific needs of, respectively, a particular language or a particular language pair.

*Proceedings of the 4th EACL Workshop on Statistical Machine Translation* , pages 90–94,
Athens, Greece, 30 March – 31 March 2009. 

### 2.1 Methodology

The translation algorithm follows the traditional pattern of a transfer system. First the input sentence is parsed by the Fips parser, producing an information-rich phrase-structure representation with associated predicate-argument representations. The parser also identifies multiword expressions such as idioms and collocations – crucial elements for a translation system (cf. Seretan & Wehrli, 2006). The transfer module maps the source-language abstract representation into the target-language representation. Given the abstract nature of this level of representation, the mapping operation is relatively simple and can be sketched as follows: recursively traverse the source-language phrase structure in the order: head, right subconstituents, left subconstituents. Lexical transfer (the mapping of a source-language lexical item with an equivalent target-language item) occurs at the head-transfer level (provided the head is not empty) and yields a target-language equivalent term often, but by no means always, of the same category. Following the projection principle used in the Fips parser, the target-language structure is projected on the basis of the lexical item which is its head. In other words, we assume that the lexical head determines a syntactic projection (or meta-projection).

Projections (ie. constituents) which have been analyzed as arguments of a predicate undergo a slightly different transfer process, since their precise target-language properties may be in part determined by the subcategorization features of the target-language predicate. To take a simple example, the direct object of the French verb *regarder* in (1a) will be transferred into English as a prepositional phrase headed by the preposition *at*, as illustrated in (2a). This information comes from the lexical database. More specifically, the French-English bilingual lexicon specifies a correspondence between the French lexeme [ $_{VP}$ regarder NP ] and the English lexeme [$_{VP}$ look [$_{PP}$ at NP ] ]. For both sentences, we also illustrate the syntactic structures as built, respectively, by the parser for the source sentence and by the translator for the target sentence.

(1)a. Paul a regardé la voiture.

b. [$_{TP}$ [$_{DP}$ Paul ] a [$_{VP}$ regardé [$_{DP}$ la [$_{NP}$ voiture ] ] ] ]

(2)a. Paul looked at the car.

b. [$_{TP}$ [$_{DP}$ Paul ] [$_{VP}$ looked [$_{PP}$ at [$_{DP}$ the [$_{NP}$ car ] ] ] ] ]

### 2.2 Adding a language to the system

Given the general model as sketched above, the addition of a language to the system requires (i) a parser and (ii) a generator. Then for each language pair for which that language is concerned, the system needs (iii) a (potentially empty) language-pair specific transfer module, and (iv) a bilingual lexical database. The first three components are described below, while the fourth will be the topic of section 3.

**Parser** The Fips multilingual parser is assumed. Adding a new language requires the following tasks: (i) grammar description in the Fips formalism, (ii) redefinition of the language-specific parsing methods to suit particular properties of the language, and (iii) creation of an appropriate lexical database for the language.

**Generator** Target-language generation is done in a largely generic fashion (as described above with the transfer and projection mechanisms). What remains specific in the generation phase is the selection of the proper morphological form of a lexical item.

**Language-pair-specific transfer** Transfer from language A to language B requires no language-pair specification if the language structures of A and B are isomorphic. Simplifying a little bit, this happens among closely related languages, such as Spanish and Italian for instance. For languages which are typologically different, the transfer module must indicate how the precise mapping is to be done.

Consider, for instance, word-order differences such as adjectives which are prenominal in English and postnominal in French – *a* **red** *car* vs. *une voiture* **rouge**. The specific English-French transfer module specifies that French adjectives, which do not bear the [+prenominal] lexical feature, correspond to right subconstituents (vs. left subconstituents) of the head noun. Other cases are more complicated, such as the V2 phenomenon in German, pronominal cliticization in Romance languages, or even the use of the *do* auxiliary in English interrogative or negative sentences. Such cases are handled by means of specific procedures,

which are in some ways reminiscent of transformation rules of the standard theory of generative grammar, ie. rules that can insert, move or even delete phrase-structure constituents (cf. Akmajian & Heny, 1975).

So far, the languages taken into account in the MulTra project are those for which the Fips parser has been well developed, that is English, French, German, Italian and Spanish. Of the 20 potential language pairs five are currently operational (English-French, French-English, German-French, German-English, Italian-French), while 6 other pairs are at various stages of development.

## 3 Multilingual lexical database

### 3.1 Overview of the lexical database

The lexical database is composed for each language of (i) a lexicon of words, containing all the inflected forms of the words of the language, (ii) a lexicon of lexemes, containing the syntactic/semantic information of the words (corresponding roughly to the entries of a classical dictionary) and (iii) a lexicon of collocations (in fact multi-word expressions including collocations and idioms). We call the lexemes and the collocations the *lexical items* of a language.

The bilingual lexical database contains the information necessary for the lexical transfer from one language to another. For storage purposes, we use a relational database management system. For each language pair, the bilingual dictionary is implemented as a relational table containing the associations between lexical items of language A and lexical items of language B. The bilingual dictionary is bi-directional, i.e. it also associates lexical items of language B with lexical items of language A. In addition to these links, the table contains transfer information such as translation context (eg. sport, finance, law, etc.), ranking of the pairs in a one-to-many correspondence, semantic descriptors (used for interactive disambiguation), argument matching for predicates (mostly for verbs). The table structures are identical for all pairs of languages.

Although the bilingual lexicon is bidirectional, it is not symmetrical. If a word $v$ from language A has only one translation $w$ in language B, it doesn't necessarily mean that $w$ has only one translation $v$. For instance the word *tongue* corresponds to French *langue*, while in the opposite direction the word *langue* has two translations, *tongue* and *language*. In this case the descriptor attribute from French to English will mention respectively "body part" and "language". Another element of asymmetry is the ranking attribute used to mark the preferred correspondences in a one-to-many translation[1]. For instance the lexicographer can mark his preference to translate *lovely* into the French word *charmant* rather than *agréable*. Of course the opposite translation direction must be considered independently.

What is challenging in this project is that it necessitates as many bilingual tables as the number of language pairs considered, i.e. $n(n-1)/2$ tables. We consider that an appropriate bilingual coverage (for general purpose translation) requires well over 60'000 correspondences per language pair.

In the framework of this project we consider 5 languages (French, English, German, Italian, Spanish). Currently, our database contains 4 bilingual dictionaries (out of the 10 needed) with the number of entries given in figure 1:

| language pair | Number of entries |
|---|---|
| English - French | 77'569 |
| German - French | 47'797 |
| French - Italian | 38'188 |
| Spanish - French | 23'696 |

Figure 1: Number of correspondences in bilingual dictionaries

Note that these 4 bilingual dictionaries were manually created by lexicographers and the quality of the entries can be considered as good.

### 3.2 Automatic generation

The importance of multilingual lexical resources in MT and, unfortunately, the lack of available multilingual lexical resources has motivated many initiatives and research work to establish collaboratively made multilingual lexicons, e.g. the Papillon project (Boitet & al. 2002) or automatically generated multilingual lexicons (see for instance Aymerish & Camelo, 2007, Gamallo, 2007).

We plan to use semi-automatic generation to build the 6 remaining dictionaries. For this purpose we will derive a bilingual lexicon by transitivity, using two existing ones. For instance, if we have bilingual correspondences for language pair

[1] This attribute takes the form of an integer between 6 (preferred) and 0 (lowest).

A → B and B → C, we can obtain A → C. We will see below how the correspondences are validated.

The idea of using a pivot language for deriving bilingual lexicons from existing ones is not new. The reader can find related approaches in (Paik & al. 2004, Ahn & Frampton 2006, Zhang & al. 2007) . The specificity of our approach is that the initial resources are manually made, i.e. non noisy, lexicons.

The derivation process goes as follows:

1. Take two bilingual tables for language pairs (A, B) and (B, C) and perform a relational equi-join. Perform a filtering based on the preference attribute to avoid combinatory explosion of the number of generated correspondences.

2. Consider as valid all the unambiguous correspondences. We consider that a generated correspondence $a \rightarrow c$ is unambiguous if for the lexical item $a$ there exists only one correspondence $a \rightarrow b$ in the bilingual lexicon (A, B) and for $b$ there exists only one correspondence $b \rightarrow c$ in (B, C). As the lexicon is non symmetrical, this process is performed twice, once for each translation direction.

3. Consider as valid all the correspondences obtained by a pivot lexical item of type collocation. We consider as very improbable that a collocation is ambiguous.

4. All other correspondences are checked in a parallel corpus, i.e. only the correspondences actually used as translations in the corpus are kept. First, the parallel corpus is tagged by the Fips tagger (Wehrli, 2007) in order to lemmatize the words. This is especially valuable for languages with rich inflection, as well as for verbs with particles. In order to check the validity of the correspondences, we count the effective occurrences of a given correspondence in a sentence-aligned parallel corpus, as well as the occurrences of each of the lexical items of the correspondence. At the end of the process, we apply the *log likelihood ratio* test to decide whether to keep or discard the correspondence.

### 3.3 Results of automatic generation

The English-German lexicon that we used in the shared translation task was generated automatically. We derived it on the basis of English-French and German-French lexicons. For the checking of the validity of the correspondences (point 4 of the process) we used the parallel corpus of the debates of the European Parliament during the period 1996 to 2001 (Koehn, 2005). Figure 2 summarizes the results of the four steps of the derivation process:

| Step | Type | Eng.-Ger. |
|---|---|---|
| 1 | Candidate corresp. | 89’022 |
| 2 | Unambiguous corresp. | 67’012 |
| 3 | Collocation pivot | 2’642 |
| 4 | Corpus checked | 2’404 |
| | Total validated corresp. | 72’058 |

Figure 2: Number of derived entries for English-German

We obtained a number of entries comparable to those of the manually built bilingual lexicons. The number of the correspondences for which a validation is necessary is 19’368 (89’022-(67’012+2’642)), of which 2’404 (approximately 12%) have been validated based on the the EuroParl corpus, as explained above. The low figure, well below our expectations, is due to the fact that the corpus we used is not large enough and is probably not representative of the general language.

Up to now, the English-German dictionary required approximately 1’400 entries to be added manually, which is less than 2% of the entire lexicon.

## 4 Conclusion

Based on a deep linguistic transfer approach and an object-oriented design, the MulTra multilingual translation system aims at developing a large number of language pairs while significantly reducing the development cost as the number of pairs grows. We have argued that the use of an abstract and relatively generic linguistic level of representation, as well as the use of an object-oriented software design play a major role in the reduction of the complexity of language-pair transfer modules. With respect to the bilingual databases, (corpus-checked) automatic derivation by transitivity has been shown to drastically reduce the amount of work.

## Acknowledgments

The research described in this paper has been supported in part by a grant from the Swiss national science foundation (no 100015-113864).

## 5 References


Ahn, K. and Frampton, M. 2006. "Automatic Generation of Translation Dictionaries Using Intermediary Languages´´ in Cross-Language knowledge Induction Workshop of the EACL 06, Trento, Italy, pp 41- 44.

Akmajian, A. and F. Heny, 1975. *An Introduction to the Principles of Generative Syntax*, MIT Press.

Arnold, D. 2000. "Why translation is difficult for computers" in H.L. Somers (ed.) *Computers and Translation : a handbook for translators*, John Benjamin.

Aymerich, J. and Camelo, H. 2007." Automatic extraction of entries for a machine translation dictionary using bitexts´´ in MT Summit XI, Copenhagen, pp. 21-27

Boitet, Ch. 2001. "Four technical and organizational keys to handle more languages and improve quality (on demand) in MT" in *Proceedings of MT-Summit VIII*, Santiago de Compostela, 18-22.

Boitet, Ch., Mangeot, M. and Sérasset, G. 2002. "The PAPILLON project: cooperatively building a multilingual lexical database to derive open source dictionaries & lexicons" in *Proceedings of the 2nd workshop on NLP and XML*, COLING 2002, Taipei, Taiwan.

Bresnan, J. 2001. *Lexical Functional Syntax*, Oxford, Blackwell.

Chomsky, N. 1995. *The Minimalist Program*, Cambridge, Mass., MIT Press.

Culicover, P. & R. Jackendoff, 2005. *Simpler Syntax*, Oxford, Oxford University Press.

Gamallo, P. 2007. "Learning Bilingual Lexicons from Comparable English and Spanish Corpora" in *Proceedings of MT Summit XI*, Copenhagen.

Hutchins, J. 2003. "Has machine translation improved?" in *Proceedings of MT-Summit IX*, New Orleans, 23-27.

Kay, M. 1997. "Machine Translation : the Disappointing Past and Present" in R.A. Cole, J. Mariani, H. Uskoreit, G. Varile, A. Zaenen and A. Zampoli *Survey of the State of the Art in Human Language Technology*, Giardini Editori.

Koehn, P. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation´´ in MT Summit 2005.

Ney, H. 2005. "One Decade of Statistical Machine Translation" in *Proceedings of MT-Summit X*, Pukhet, Thailand.

Paik, K., Shirai, S. and Nakaiwa, H. 2004. "Automatic Construction of a Transfer Dictionary Considering Directionality", in COLING 2004 Multilingual Linguistic Resources Workshop, Geneva, pp. 25-32.

Seretan, V. & E. Wehrli, 2006. "Accurate Collocation Extraction Using a Multilingual Parser" in *Proceedings of the ACL*, 953-960, Sydney, Australia.

Wehrli, E. 2007. "Fips, a "deep" linguistic multilingual parse" in *Proceedings of the ACL 2007 Workshop on Deep Linguistic processing*, 120-127, Prague, Czech Republic.

Zhang, Y., Ma, Q. and Isahara, H. 2007. "Building Japanese-Chinese Translation Dictionary Based on EDR Japanese-English Bilingual Dictionary" in *MT Summit XI*, Copenhagen, pp 551-557.