Systran System

by

S. Traboulsi

Gentlemen:

We shall speak about some particularities and difficulties we encountered in translation into Arabic.

As it was mentioned before, transfer problems increase with the distance between source and target languages, and Arabic is very different from Europan languages. So the translation will be very difficult, especially if we want to have a good concise style in Arabic.

– First I shall mention the inversion of verb/subject order. It may seem very easy like in the sentence,

THE CHILD WOKE UP. استيقظ الطفل

but it often happens that the subject is very far from the verb and causes a bad translation like in the sentence.

THE COMPANY, WHICH WAS FOUNDED SEVERAL YEARS AGO AND (...) IS NOW BEGINNING TO PRODUCE TELEVISIONS

تبدأ الآن الشركة، التي أسست منذ سنوات عديدة و. . . ، بإنتاج الأجهزة التلفزيونية

the correct translation may necessitate the use of a nominal sentence.

إن الشركة التي. . . قد بدأت بإنتاج الأجهزة التلفزيونية

– But we have also many reordering problems like the obligation to insert the subject or object between the verb and the expression complementing the verb. We solved this problem at the level of the dictionary by introducing a special code indicating this kind of reorder as in the translations.

TO REFUEL A PLANE زوّد (الطائرة) بالوقود

TO SUBLET A HOUSE أجّر (المنزل) باطنياً

– The most important problem of this order is when you have a complete chain of noun complements and adjectives.

It is really easy to translate

THE HOUSE OF THE MAN or	منزل الرجلِ
THE RED HOUSE	المنزلُ الأحمرُ

and not so tricky to translate

THE RED HOUSE OF THE MAN	منزل الرجلِ الأحمر

even if it is a little ambiguous.

But the situation becomes really complicated with several complements and adjectives, not because of the multiple applications of the rule, but because the output sentence begins to be very ambiguous. Human translators use a little trick by separating sometimes a particular noun complement by an expression like (الخاص ب ، التابع لـ) equivalent to the English (of ...) and that is function of the ambiguity of attributing particular adjectives to the right noun.

Such work is very difficult, if not impossible, to achieve automatically, because we would need a complete understanding of the universe. Here I mean to know the exact set of adjectives that can describe each noun.

– The Arabic language prefers also direct forms.

THE HOUSE WAS CONSTRUCTED BY THE COMPANY – THE COMPANY CONSTRUCTED THE HOUSE –	بَنَت الشركةُ المنزلَ

but systematic transformation of these structures may lead to bad situations like:

THE CHILD WAS TOLD BY HIS FATHER TO STAY QUIET.

طلب أبوه من الولد البقاء هادئًا

Using direct form gives

طلب الأب من ولده البقاء هادئًا

But detection of reversal relationships is very difficult, if not impossible. So the best thing to do is to transform passive forms into active forms only when a series of conditions is realized.

– Pronouns are also a classical problem that can have some tricks.
For instance it is easy to translate :

THE MAN ASKED HIM ABOUT SOMETHING.	سأله الرجل عن شيء ما

But if you don't take precautions, you will translate

HE ASKED HIM AND HER ABOUT SOMETHING

by سأله وها عن شيء ما

instead of سأله وإياها عن شيء ما

– Another important problem is the modification of the nature of words during the translation. For instance, an adverb is transformed into a noun in Arabic, and then you are obliged to modify a complete set of relationships inside the sentence to achieve good inflexion. The number problem has also the particularity of necessity of according inflexion with the grammatical role in the sentence and to accord the gender with the noun described by the number

اشتريت سبعةَ أقلامٍ

– We dealt also with the dual problem, and it is not easy to determine wether the subject of a particular verb is dual or plural. We refer generally to constructions like:

THEY BOTH WANT TO DRINK COFFEE.

or

THE TWO MEN DRINK COFFEE.

but also we must track the pronouns in the following sentences to continue to use the dual form like in the sentence.

THE TWO MEN WALKED IN THE STREET. THEN, THEY DRANK COFFEE.

Of course, we can trap the system with constructions like:

THE TWO MEN WALKED IN THE STREET. THEY WERE JOINED BY ANOTHER PERSON, AND THEY DRANK COFFEE.

It is not easy to explain to the machine that now they are three. Of course, this kind of situation seldom happens in the texts we translate.

– I will also mention the style problem of the continuation prepositions أحرف الاستئناف that must be inserted in the Arabic text. The problem of the beginning of sentences is a little harder. You will translate for instance;

WHILE HE IS WRITING

by وهو يكتب
but if you have in the beginning

HE IS WRITING.

you will say: إنه يكتب

or if you have (...) and he went to the movies وذهب إلى السينما
but
HE WENT TO THE MOVIES.

will be translated by. لقد ذهب إلى السينما

– There are also all the particular constructions that must be dealt with, like age structures, or date or hour structures and you shall not give those translation:

How old are you? كيف تكون قديماً أنت

On December 1st على ديسمبر الأول

At half past ten عند نصف بعد عشرة

and here you need lexical routines that are very numerous in the English-Arabic Translating System.

– But the real problems emerge only from translations that require knowledge of the universe of the text to be translated.

Earlier I mentioned the problem of dividing a series of noun complements, but there are also other very good linguistic subjects as translation of impersonal forms followed by (to have) like:

IT IS POSSIBLE TO HAVE A PEN BY BUYING IT

ـ من الممكن الحصول على قلم بشرائه

IT IS POSSIBLE TO HAVE SOME IDEAS

من الممكن ملك بعض الأفكار

but those impersonal forms followed by (to be) are far more difficult.

IT IS POSSIBLE TO BE A GOOD STUDENT.

The translation may be then. من الممكن أن تكون / يكون؟ طالباً مجداً

It requires knowledge about whom we are talking about. And that is not easy at all.

Let us now consider the dictionaries.

Like other SYSTRAN systems, the English Arabic system has two main kinds of dictionaries.

The basic stem dictionary gives for each source term a complete morphological, grammatical, syntactical and semantical description, and gives a basic translation of this term with the morphological, and grammatical description in the target language.

All those items are normal features. But there are also many functional descriptions that help the synthesis programs. For instance, for the proper place where subject or object must be inserted with respect to an adverb or a preposition, or the inflexions it gives when, for instance, a transitive English verb is translated by an instransitive Arabic verb.

But we also have very useful functions that allows us to transform an adjective into a verbal clause, the subject or the object of which is a pronoun referring to the noun described by the adjective.

You can for instance translate:

UNBEARABLE by = (لا يطاق / لا تطاق)

INCORRIGIBLE by = (لا يمكن تصحيحه / لا يمكن تصحيحها)

All those functions give a lot of flexibility in choosing terminology.

I would like to insist on the importance of a good choice of the meaning for polysemous terms. Practical experience showed us that generally it is not the most largely used meaning that must be coded, but the most general term to avoid some bad translations.

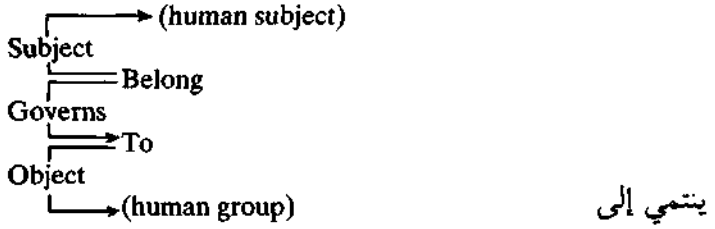
But most important is the contextual dictionary which will allow us to make

homographic choices, to change grammatical results or to choose a particular meaning depending on the context of a word.

In the beginning, SYSTRAN was using what we called "Topical Glossaries", and the translator had to indicate the field of the translation to the machine to obtain the wanted meaning. For instance "plant" was translated by مصنع in general, and by نبتة for agricultural use.

But it did not work, not only because it was a constraint for the translator, but because experience showed that even in agricultural texts the meaning مصنع was more frequent than نبتة . Besides, it was impossible to process نبات inside مصانع for instance.

So, a contextual dictionary is a very useful tool that allows: real translation without intervention of the translator. And then if you translate the verb "to belong" in the stem dictionary by تابع لـ or ملك you may introduce an expression where you will say that.



and you will translate

THIS PLAYER BELONGS TO CAIRO BLUB

ينتمي هذا اللاعب إلى نادي القاهرة

As may be seen, we can use in those expressions both syntactic relationships and semanti properties.

Such expressions are little programs and you must write a little program for each meaning, and the contextual dictionary is a compliation of thousands and thousands of those programs.

This feature is unique in SYSTRAN, but is absolutely necessary when languages are very far from each other. It seldom happens in this case that an ambiguous term may be translated by an equivalent ambiguous meaning.

Besides, in English-Arabic systems, many terms must be translated in different manners depending on the sentence, like, for instance, reflexive verbs:
to move something: نقل شيئاً to wash something: غسل شيئاً

to move: انتقل To wash (oneself): اغتسل

Of course, you cannot write such programs as if you were putting a simple meaning as a translation to a simple entry, and those programs must be written by well trained lexicographers.

Besides, experience showed us that we must be very careful when implementing the dictionaries, because those little programs are very powerful and can influence all the words in the sentence. We must absolutely avoid unrealistic situations. The choice to implement a particular meaning must emerge from day to day use of the system and the situation must be dealt with on the basis or real use of the system. It is not realistic to code an exceptional or artificial meaning which occurs once in several millions of pages, if it introduces a probability of error of one thousandth on the basic meaning. This is why the implementation of dictionaries must be done by a very experienced team.

To answer the questions that will be surely asked, our intent is absolutely not to impose particular terminology. The fundamental work is on the determination of the source meaning. Arabic terms may be chosen at any time by the customer for his system, and may be changed very easily.

The implementation of dictionaries is a very important work. It is very easy to introduce the first thousands of terms but the work begins to be really difficult for the fourth meaning of the hundred thousandth term.

Besides, it is very important to note that SYSTRAN dictionaries may gather nearly an infinite set of terms, depending on the number of discs.

Now we have more than a hundred thousand entries in our dictionaries, but we will have approximately 150.000 by the end of June taking advantages of the work achieved by SYSTRAN Institute in European pairs and we think that it is a minimum for a general dictionary.

I shall say some words now about text treatment.

Automatic Translation process involves several operations, and each operation must be done in a cost effective way, from inputting until outputting.

– At the input we connected a self training scanner whch can read 16 fonts and can be taught others.

– Errors in the source text were of course unintelligible for the computer, so we also adopted a self detecting errors program to point to each word which is not in the dictionary.

– The SYSTRAN System is of course implemented on an IBM computer allowing very fast translation at a speed of 300.000 words an hour.

– The translation may be then post-edited on a terminal, the upper window of which contains the source text and the lower on the Arabic text.

Therafter, the text may be printed on a laser printer or directly composed on very high quality electro-erosion composer, allowing the composition of the majority of texts including all schemes.

We are testing such an organisation by using it in a day to day work in Paris in the English French and French English pairs with the help of the French National Union of Translation Companies to detect all negative points that may exist in it.

As you can see, we are always following a very programed way in developing all aspects of machine translation. But we also worry about the theoretical aspects, and not only we study all developments of SYSTRAN System, but we are also very interested in several developments done by univesities. We are now establishing agreements with several universities in France like LYON University and NANCY University for studies concerning such matters as semantic construction, Arabic morphology generation and Arabic voice synthesizing.

And the modular structure of SYSTRAN will allow the integration of more improvements all along the time.

* * *