French to Arabic Machine Translation: the Specificity of Language Couples

Haytham Alsharaf [1], Sylviane Cardey[2] and Peter Greenfield[3]

[1]French Language Unit, [2,3]Centre Tesnière
[1]Kuwait University, [2,3]Université de Franche-Comté, France
[1]alsharaf2001@yahoo.com, [2]sylviane.cardey@univ-fcomte.fr, [3]peter.greenfield@univ-fcomte.fr

Abstract. We present an approach to the machine translation of French to Arabic in which we explain each of the different steps and some of the problems encountered at each level of linguistic analysis together with their solutions. We show that for high quality translation, an approach which is both global and proper to the couple French to Arabic is necessary, and we compare our approach with existing ones (direct, transfer, pivot, statistical). The fact that we have two languages which are linguistically distant requires that certain linguistic phenomena that are specific to the couple must be analysed, such phenomena not necessarily being found in other language couples.

1. Introduction

Approaches to machine translation from French to Arabic are rare. Certainly, translation memory systems do exist (as for example the commercial system An-Nakel Al-Arabi from French to Arabic from CIMOS (Paris)[1]) but the limitations of such systems are now well known to all specialists working in the domain. In this paper we present an approach which we have implemented specially for the machine translation of specialty languages from French to Arabic.

At the outset of the project, existing approaches (see for example [1], [2]) were examined (direct, transfer, pivot, statistical) but it became apparent that each presented limits and it became increasingly evident that it was necessary to devise an approach specific to this particular language couple. The fact that we have here two languages that are very distant linguistically means that, from the linguistic point of view, certain linguistic operations which are specific to this couple will not necessarily be found in other couples, whether or not one or other of French or Arabic is present as either source or target. The approach that we describe is thus unique for the couple French to Arabic; and it follows that it is not valid for the inverse couple Arabic to French. Reversing the translation steps will not result in a machine translation system of the couple Arabic to French because to have such a system would necessitate certain linguistic operations that are not present for French to Arabic. Indeed, this consideration of direction in translation explains our use of the term "language couple" rather than "language pair", by analogy with "couple" in statics in which direction is inherent. Other work that we have carried out vindicates our findings concerning the uniqueness of couples for linguistically distant language machine translation (for example Chinese to French [3], Korean to French [4], both Chinese to French and French to Arabic [5]). It is our opinion that if machine translation is to succeed, and by success we mean quality machine translation, it is necessary to limit one's ambitions by dealing with one couple of languages at a time, this implying in one given direction, and above all when the languages are linguistically distant. Moreover it is necessary to limit the domain to be translated, as for example in [6] and [7]. This means that the domain itself must be analysed linguistically in order to know and be able to formalise its morpho-syntactic, lexical and semantic behaviour. As well as this, there is the need to understand the domain's behaviour and its conceptual organisation. The principal domain we have researched is the specialty language of French and Arab law.

To conclude this introduction, we are of the view that machine translation approaches ought to treat all the linguistic levels as well as the problems that ensue in order that the results be acceptable in terms of quality. In machine translation, the same defects

[1] http://www.cimos.com/index.asp?src=fiche
"Highlights:
...
- Alignement de mémoire et de phrase de traduction"

keep being reported (see [8]). We certainly do not claim to have solved all the problems; rather we simply say that we have started the process for the machine translation, where this is possible, for French to Arabic. In the paper, we present the steps which allow the machine to perform the translation in presenting the problems encountered and proposing, where possible, solutions.

2. The translation process

Our analysis is carried out in several steps; this is due to the fact that we have two languages of different origins, and this renders the machine translation process more complex.

2.1 The steps of the analysis

Step 1: Segmentation

The user having entered an utterance in French, this step is concerned in segmenting this into French linguistic units.

Step 2: Morphological analysis

This step has as objective the production of a structure in which the French linguistic units are tagged in an unambiguous manner with their respective grammatical categories (parts of speech); that is, there are no ambiguities due to the French units' form. This restrains the possibility of ambiguity (without always excluding it completely). In the event that an utterance has more than one interpretation, the system could choose the most frequent, or preferably, the morphological analysis ought to be completed with a semantic analysis.

Step 3: French linguistic units' translation

This is the first step in which the two languages are put into relation with each other, because during this step, for each source language French linguistic unit, the system provides all of the possible Arabic translations. In the case of for example Arabic verbs and adjectives, the forms for the two genders (masculine/feminine) and the two numbers (singular/plural) are mentioned in order to enable the resolution of problems arising from differences in gender and number between the two languages.

From and including this step we prefer to use the apparently (but incorrectly) more general term linguistic unit; what started as a French linguistic unit takes on other aspects as it proceeds through the translation steps of the process. From now on a "French linguistic unit" refers to the French aspects of a given linguistic unit.

It is during this step that a part of the problem concerning the translation of prepositions is resolved. The difficulty is that the sense of prepositions and thus their translation is very dependent on their context. Sometimes indeed, a French preposition has no translation in Arabic. For example the French preposition 'à' can have up to eight possible translations in Arabic, and this means that the formal rules which select the good translation have to be found. French preposition processing for translation to Arabic cannot be done in one single step; in some situations one has to wait until the 6th step for a preposition to have its translation calculated. Thus during this 3rd step the system verifies if the lexical unit could have a preposition and if this is the case one provides at this point the translation of the actual preposition. By default, the preposition has no translation in Arabic. Even if this is in fact the case, the system cannot and does not exclude the possibility of giving a translation of the preposition as a linguistic unit later in the translation process, after the application of formal rules in the 5th step. The following example shows this type of problem where 'des' is not translated during this 3rd step but is so during the 5th:

l'ordre juridique interne **des** Etats
النظام القانوني الداخلي ل الدول

Another problem that can occur in processing prepositions is when a preposition that occurs more than once in a sentence keeps a particular sense for at least two of its occurrences and where this sense is not present in the system's lexicon for linguistic units governed by the preposition, this being because in fact the correct interpretation is done at the syntactical level. Consider the following example:

l'ordre administratif connaît *des* litiges relatifs **à** l'organisation et **au** fonctionnement *des* services publics et **aux** contrats administratifs

Here, 'à' (present in 'au' and 'aux') keeps its translation (ب) established for 'à l'organisation ' for the rest of the sentence:

au fonctionnement (بألية) **aux** contrats (بالعقود)

In consequence we have written a formal rule allowing the resolution of this type of problem. It should also be remarked that whilst the first 'des' in the above French sentence has a precise translation (في), this is not so for the second 'des'. These examples indicate the level of complexity of the

formalisation of the problem and also the need to formulate precise rules which are both complete and also which operate in concert. In respect of this we have been able to cope with these types of problem and obtain good results as is demonstrated by the complete translation of the above sentence given later in the paper.

Another problem that is treated in this 3rd step is that of the polysemy of certain lexical units, even in the case of the machine translation of specialty languages. Our definitions of simple and compound linguistic units enable a partial resolution of this problem. However, in certain cases, we find total ambiguity and this leaves us with no choice other than to interrogate the end-user in order to select a given translation. We have observed the same problem in other language couples [3], [4]. For example, the word 'direction' has different translations in the following two sentences which are otherwise similar:

le pouvoir de direction (الادارة) doit être exercé

le texte institutionnel permet de préparer les directions (التوجهات) et règlements d'application

In the steps that follow, we treat the linguistic units as members of a syntagmatic structure forming an utterance. At this point in the translation process one can ask the following questions:

- How can one transpose a French utterance into the equivalent in Arabic?
- How can one take account of particularities both in the French utterance and in the Arabic one?
- In what way does one establish the formal link between two languages which have different morpho-syntactic structures?

In order to reply to these questions, we have designed and implemented 3 successive steps (4, 5 and 6) which establish a correct morpho-syntactic and semantic transition from French to Arabic, the final lexical generation into Arabic being done in the 7th step.

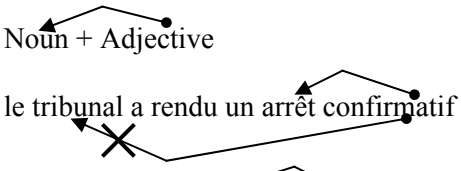
Step 4: Linguistic unit actualisation and morpho-syntactic and semantic agreement

The 4th step in the machine translation process analysis is both important and necessary in terms of the translation process formalisation.

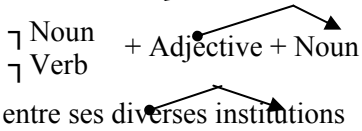
Having in the 3rd step noted all the possible translations of each linguistic unit, in this 4th step the system first actualises the following information for each linguistic unit: grammatical category, gender, number, potential prepositions, semantic information, and whether or not the linguistic unit takes the Arabic article 'al-'.

The system then performs a morpho-syntactic and semantic agreement analysis between the different linguistic units of the utterance. The object of this analysis is to establish the agreements in gender and number between the relevant linguistic units and to select one translation for each linguistic unit amongst the possibly several proposed by the previous 3rd step. In Arabic, because verbs as well as adjectives have gender and number, several translations are possible. This problem, which would appear simple to resolve at first sight, is in fact very demanding. Not only do all the morpho-syntactic dependency relations between the different grammatical categories of Arabic sentences have to be described in a formal and subsequently machine interpretable manner, but selecting the gender and number of verbs and adjectives implies automatic recognition of the sentences' subject. It seems that we have obtained good results for this analysis. In the following examples, the arrows indicate the gender dependency of one linguistic unit to another (here we are dealing with the Arabic aspect of linguistic units). We show only a few examples without entering into the details; this problem alone warrants an article in its own right in order to show all the subtleties concerning the formalisation and the processing.

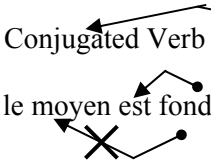
i. Noun + Adjective



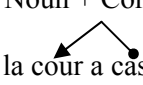
ii. ┐Noun / ┐Verb + Adjective + Noun



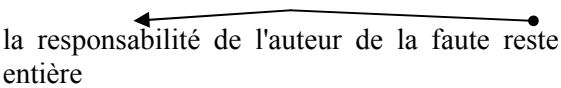
iii. Conjugated Verb + Adjective



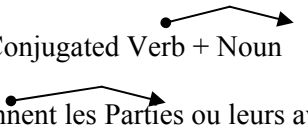
iv. Noun + Conjugated Verb



v. Noun + Noun + Noun + Conjugated Verb



vi. No Noun + Conjugated Verb + Noun



Puis interviennent les Parties ou leurs avocats

In this 4th step the polysemy of certain French grammatical linguistic units such as 'par' is dealt with. The system's lexicon contains information indicating that if a given verb is followed by 'par', then which sense is applicable; for example for the verb 'constituer', we indicate its meaning when it is followed by the preposition 'par'. 'par' can have at least four possible translations according to the context. Thus this step also involves semantic processing.

In respect of semantic processing and disambiguation, for nouns the system's lexicon indicates by means of semantic features the instances of nouns which are members of conceptual sets. Thus a 'chambre' is a part of a 'tribunal' which in turn belongs to a 'jurisdiction' and so on. This has a bearing on the translation of prepositions. Consider the following example:

la chambre criminelle de la Cour de cassation de Lyon

where the first and third 'de' acquire a precise translation (ل) due to the conceptual set membership of the noun that it is linked to.

We have classified the nouns in three categories: "specific" being for basic elements in the juridical organisation ('chambre), "generic" as in 'Cour de cassation' and finally "absolute" such as the names of towns ('Lyon') or of countries. The categories are indicated by the symbols +, ++ and +++ respectively. In steps 5 and 6 which follow, there are formal rules involving symbol processing which insert the relevant Arabic preposition between the nouns.

This 4th step is also concerned with whether, in principle, a linguistic unit can have the Arabic article '–al'. It is by means of the symbol processing that the system decides between inserting, suppressing, or keeping the article.

Steps 5, 6 and 7

The 5th step is not concerned with morpho-syntactic and semantic agreement; instead the system keeps in memory the results of operations performed in the preceding steps concerning the linguistic units in terms of selections of gender and number together with certain features. The result of these preceding operations is a symbolic structure which incorporates aspects of both the source language French and the target language Arabic.

The 5th step's symbolic structure is the basis for the generation, in the 6th step, of another symbolic structure, this of the target language. Formal rules are involved in the transfer from the 5th to the 6th step; these rules are concerned with variously suppressing, inserting or repositioning certain of the symbols that represent the linguistic units in order to have an Arabic sentence which is both syntactically and semantically correct. Thus what is involved here is not simply putting into correspondence source language and target language structures. Finally, in the 7th step, the lexical generation is achieved by replacing the symbols constituting the structure generated by the 6th step by the appropriate Arabic linguistic units.

We mention that the system's lexicon has been designed to function with the steps of the analysis. This shows once again that for machine translation systems, even the system's lexicon ought to be designed in terms of the way the system functions. It is thus preferable firstly to define the functionality of the system, from this create a model, and only then design the lexicon in function with the model.

2.2 Example of translation step by step

The following example shows the different steps of the analysis.

1 **l'ordre administratif connaît des litiges relatifs à l'organisation et au fonctionnement des services publics et aux contrats administratifs**

2 n adj vconj prep n adj prep n coord prep n prep n adj coord prep n adj

3 ن ب ل يحكم تحكم نظام اداري ادارية مرتبطين مرتبطات مرتبطة الى نزاعات ب على في التنظيم و
ل آلية ل الى على من ب خدمات عاميين عامة و ب
ل على الى عقود اداريين اداريةل

4A N(m-s-o) **adj(m-s) v.conj(m-s)** prép N(f-p-o) **adj(f-s)** prép N(m-s-x) coord **adj(f-s) v.conj (f-s) adj (m-s)**

prep N(f-s-x) X N(f-p-o) **adj(f-s)** coord prép N(f-p-o).

adj (m-s).

4B N(m-s-o) adj(m-s) v.conj(m-s) prép N(f-p-o) adj(f-s) prép N(m-s-x) coord prep N(f-s-x) X N(f-p-o) adj(f-s) coord prép N(f-p-o).

5 No adj vconj prep No adj prep Nx coord prep Nx X No1 adj coord prep art No art adj

6 No1 Art Nx prep coord Nx prep adj Art No Art prep vconj adj Art No Art adj Art No Art prep coord

7 ال نظام ال اداري يحكم في ال نزاعات ال مرتبطات ب التنظيم و آلية ال خدمات ال عامة و ب ال عقود ال ادارية

3. The global approach

In this section we wish to show the originality of our approach, why we call it a "global" approach, and furthermore we compare it with existing approaches.

First of all, we wish to make clear that our approach is not one in which the languages are analysed separately. From and including the 3rd step the analysis involves both languages; the translation is done progressively as the analysis advances.

3.1 Comparison with the direct approach

As in the direct approach, it is also difficult to add other languages; both approaches have need of extensive lexica. However, our approach does have abstract representations and a set of transformational rules which manage the transfer from one representation to another. As well as this, for a given linguistic unit our system provides all the possible translations; in no case does our system stock a single sense for each word.

3.2 Comparison with the transfer approach

The idea of abstract representations exists just as well in our approach as in the transfer approach. In the latter there is a distinction between the source language abstract representations and those of the target language; the two are distinct, which is not the case in our approach. Furthermore, after the morphological analysis, our approach provides all the possible translations of each linguistic unit, whilst in the case of the transfer approach no such translation is provided before the completion of the analysis; this could be reasonable in the case where there is a bijection between forms and senses. The transfer approach could translate "near" languages and where the domains are limited. However it cannot be used in our translation of French to Arabic where the final translation is not the result of some transfer link but of a morphological, syntactic and semantic "construction".

3.3 Comparison with the pivot approach

In our system as in the pivot approach a linguistic unit can have several senses and several translations, the sense which is retained is that generated by calculations on the context. This aspect is the only feature which is common between the pivot approach and ours.

3.4 Comparison with the statistical approach

As is the case in the statistical approach, our system chooses a given morphological structure based on frequency data, but this occurs only if there are several possible structures. However recourse to this option, whilst certainly useful, is risky; furthermore this occurs only in step 2.

4. Some results

To give an idea concerning the performance of our system, we give here a set of examples.

- une promesse de vente a été conclue entre Max et Luc
  وعد ال بيع أبرم بين ماكس و لوك
- l'objet de la demande principale est l'annulation de jugement et rejet de toutes les demandes présentées par Max à titre subsidiaire
  موضوع ال طلب ال رئيسي يكون الغاء ال حكم و رفض كل ال طلبات ال مقدمة منقبل ماكس بعنوان احتياطي
- le pouvoir de direction doit être exercé localement et conformément aux lois locales
  سلطة ال ادارة تجب أن تكون معمول بها محليا و وفقا ل ال قوانين ال محلية
- la qualification retenue par la Cour de cassation dans son arrêt doit être approuvée
  ال تكييف ال محسوم منقبل محكمة التمييز في حكم ـه يجب أن يكون مقبول
- le Tribunal de loi martiale condamne à des lourdes peines les détenues
  محكمة القانون العرفي تدين ب عقوبات شديدة ال معتقلين
- il relève de la compétence administrative tout ce qui concerne l'organisation et le fonctionnement des services publics
  يكون من ال اختصاص ال اداري كل ما يتعلق ب التنظيم و آلية ال خدمات ال عامة

- Luc a quitté la France avant l'examen de son pourvoi par la Cour de cassation
  لوك غادر فرنسا قبل فحص التماس ـه منقبل محكمة التمييز
- la responsabilité de l'auteur de la faute reste entière
  مسئولية ال فاعل ال خطأ تبقى كاملة
- les avocats réclament la libération de leur client devant la Chambre d'accusation
  ال محامين يطالبون ال افراج عن موكل هم أمام غرفة لاتهام
- Max a été pris en flagrant délit de vol dans un grand magasin
  ماكس قبض عليه متلبسا بجريمة ال سرقة في محل كبير
- d'après la loi française , il est interdit de faire travailler les femmes la nuit
  حسب ال قانون ال فرنسي يحظر عمل ال نساء ال ليل
- la directive stipule que les hommes et les femmes doivent avoir les mêmes droits dans leurs activités professionnelles
  ال مرسوم يوضح أن ال رجال و ال نساء ملزمون أن يملكون ال نفس حقوق في نشاطات هم ال مهنية
- attendu que nul n'est censé ignorer la loi , tout citoyen est assujetti au droit commun
  بما أن لاعذر لاحد بجهل القانون كل مواطن يكون خاضع ل ال قانون ال مشترك
- la constitution française décrit l'organisation de l'Etat et définit les règles du jeu entre ses diverses institutions
  ال دستور ال فرنسي يصف التنظيم الدولة و يحدد ال نظام بين مؤسسات ـه ال متعددة
- le président de la République peut , après consultation du Premier ministre et des présidents des assemblées , prononcer la dissolution de l'Assemblée nationale
  رئيس ال جمهورية يستطيع بعد استشارة ال رئيس الوزراء و رؤساء ال مجالس أن يعلن حل الجمعية الوطنية
- tout ressortissant de la Communauté peut demander au juge de son pays l'application des traité directives et décisions communautaires
  كل تابع ل ال أتحاد يستطيع أن يطلب من قاضي بلد ـه تطبيق توجهات و قرارات اتحادية

5. Conclusion

What we have endeavoured to show in this paper is that in our view French cannot be translated into Arabic with systems based on existing approaches (direct, transfer, pivot, statistical). This does not mean that our approach does not draw on aspects of these existing approaches; on the contrary it does incorporate certain aspects of all of them in certain of our system's steps. Indeed, this grouping of the operations (and there are many more than have been described here) in the form of steps (which comprise sub-steps which are not developed in the paper) is so that the presentation is clearer and also so as to avoid entering into the details of each problem; the grouping in steps is in fact a simplified view of the translation process. The system carries out calculations which insert or suppress particular units (such as '–al') without necessarily attaching them to other units at that moment; the attaching can be done later in the translation process. What is to be noted is that the overall structure of our system is different from these existing approaches and that we also use functions not found in the existing approaches, functions which are indispensable for treating the type of language couple characterised by having linguistically distant languages.

Whilst the approach that we have devised is specific to the couple French to Arabic, the approach itself, as far as this specific couple is concerned, is global in nature. It is in coming to terms with this apparent contradiction that we feel that machine translation at least for linguistically distant languages can advance, that is in applying such a global approach (to a specific language couple) in a global manner (to arbitrary language couples), an approach which might be called "global^2". By global in terms of a specific language couple we mean that although we organise the translation process in steps (there is never backtracking over steps), the results of a given step are not all simply processed as a whole "immediately" by the following step; the various parts of the results are processed by succeeding steps that have interest in doing so. This global approach to a specific language couple is a reflection of the impossibility of analysing language in terms of separate "levels" (lexis, morphology, syntax, semantics etc.) [9].

Finally, we also wish to stress the fact that the types of information required by the various methods that are used in machine translation are not the same, and that it is only when a translation

model has been created for a given machine translation system that one can search for the information types and the data necessary for completing the system's development. This is what we have done for a part of the domain of law, and also for another domain which is very much simpler, that of weather forecasts, as illustrated by the following example:

ailleurs sur la moitié nord-ouest du pays, la matinée sera peu nuageuse et brumeuse sur le sud-ouest

الىموضعأخر على ال قسم ال شمال-غربي ل ال

بلد ال صبحية ستكون غائمة قليلا و معتمة على ال جنوب-غرب

References

[1] Hutchins, W.J. and Somers, H.L., *An Introduction to Machine Translation*, Academic Press, London, 1992.

[2] Clas, A., Bouillon, P., *La Traductique*, Les presses de l'université de Montréal, AUPELF-UREF, Montreal, 1993.

[3] Shen, Y., Cardey, S., "Vers un traitement du groupe nominal dans la traduction automatique chinois-français", *Ve Congrès International de Traduction,* Barcelona, 29-31 October 2001.

[4] Cardey, S., Greenfield, P., Hong, M-S., "The TACT machine translation system: problems and solutions for the pair Korean – French", *Translation Quarterly*, No. 27, The Hong Kong Translation Society, Hong Kong, 2003, pp. 22-44.

[5] Alsharaf, H., Cardey, S., Greenfield, P., Shen, Y. "Problems and Solutions in Machine Translation Involving Arabic, Chinese and French", Paper accepted for publication by the IEEE Computer Society Press in the proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004) to be held 5-7 April 2004, in Las Vegas, Nevada, USA.

[6] Aizawa, T. *et al*. "A Machine Translation System for Foreign News in Satellite Broadcasting", *COLING-90*, 1990.

[7] Isabelle, P. *et al.* "TAUM-Aviation : description d'un système de traduction automatisée des manuels d'entretien en aéronautique", *COLING-78*, 1978.

[8] Hutchins, J., "Has machine translation improved? some historical comparisons", Proceedings of the Ninth Machine Translation Summit, New Orleans, USA, September 23-27,2003, pp.181-188.

[9] Cardey, S., Greenfield, P., "Peut-on séparer lexique, syntaxe, sémantique en traitement automatique des langues ?", Cahiers de lexicologie 71 1997-2, pages 37-51, ISSN 007-9871.