Building a Parallel Multilingual Corpus (Arabic-Spanish-English)

Doaa Samy[*], Antonio Moreno Sandoval[*], José M. Guirao[†], Enrique Alfonseca[‡]

[*] Computational Linguistics Laboratory, Universidad Autónoma de Madrid
[†] Department of Computer Languages and Systems, Universidad de Granada
[‡] Department of Computer Science, Universidad Autónoma de Madrid
[*]{doaa, sandoval@maria.lllf.uam.es}, [†] jmguirao@ugr.es, [‡] Enrique.Alfonseca@uam.es

Abstract

This paper presents the results (1st phase) of the on-going research in the Computational Linguistics Laboratory at Autónoma University of Madrid (LLI-UAM) aiming at the development of a multi-lingual parallel corpus (Arabic-Spanish-English) aligned on the sentence level and tagged on the POS level. A multilingual parallel corpus which brings together Arabic, Spanish and English is a new resource for the NLP community that completes the present panorama of parallel corpora. In the first part of this study, we introduce the novelty of our approach and the challenges encountered to create such a corpus. This introductory part highlights the main features of the corpus and the criteria applied during the selection process. The second part focuses on two main stages: basic processing (tokenization and segmentation) and alignment. Methodology of alignment is explained in detail and results obtained in the three different linguistic pairs are compared. POS tagging and tools used in this stage are discussed in the third part. The final output is available in two versions: the non-aligned version and the aligned one. The latter adopts the TMX (Translation Memory Exchange) standard format. At the end, the section dedicated to the future work points out the key stages concerned with extending the corpus and the studies that can benefit, directly or indirectly, from such a resource.

1. The LLI-UAM Multilingual Parallel Corpus: A New Resource[*]

1.1. State-of-the-art

Much work has been carried out in the field of developing parallel corpora either bilingual or multilingual. However, in our opinion, there are two main reasons behind the uniqueness and novelty of our corpus. Both reasons are directly related to the state-of-the-art in the field.

First, there is a significant gap between the number of resources available for English and Spanish, on one hand, and the resources available for Arabic, on the other hand. This unbalance is reflected on the studies concerning parallel corpora and especially those dealing with Arabic. In most of the cases, they are bilingual studies in combination with English. The results of the survey we conducted to locate Arabic parallel corpora prove this fact. There are the four corpora available through the LDC:

1. UN Arabic English Parallel Text (LDC2004E13)
2. Arabic News Translation Text Part 1 (LDC2004T17)
3. Multiple Translation Arabic (MTA) Part 1 (LDC2003T18)
4. Arabic English Parallel News Part 1 (LDC2004T18)

Second, major initiatives aiming at developing multilingual corpora have been taken within the framework of various European projects such as CRATER (Garside et al. 1994), MULTEXT (Ide & Veronis 1994), and ECI/MCI. More recent are the initiatives of OPUS (Tiedemann & Nygaard 2004) and EUROPARL (Koehn 2005). Therefore, the coverage is limited to the European languages and Arabic language is not included.

Taking into consideration both factors, we insist on the fact that the corpus, we are presenting here, is the first parallel corpus offering the following language combination (Arabic-Spanish-English).

1.2. Building the corpus

The selection process was characterized by a number of challenges and difficulties to meet the established criteria in terms of quality and quantity. Finding a considerable quantity of quality texts available in the three languages was our main endeavor. The quality in this case is directly related to the nature, source and the translation of the selected texts. Representativeness, availability in electronic format and legal use are other relevant issues in this stage.

To apply these criteria, the following decisions were taken:

1. Texts should not be automatically translated.
2. Texts should represent the modern standard use of the language
3. Sources should be freely available in electronic format.
4. Author's copyrights should be respected and the use of the text should be within the principle of Fair Use.

Opting for the United Nations documents was the most practical and feasible solution. The reasons behind could be summarized in the following:

1. Arabic, Spanish and English are among the official languages of the Organization.
2. Translation quality is guaranteed.
3. Texts represents a modern standard use of the language.
4. Documents are available freely and in considerable quantities.
5. UN explicitly states that using texts for academic purposes is considered a “fair-use”.

1.3. Basic Features

In this first stage of the research, the total size of the corpus is about 3 million words divided into three main

[*] Part of this research has been supported by the grant TIN2004-07588-C03-02 (Spanish Ministry of Education and Science).

subgroups corresponding to the three languages Arabic (901,511 words), Spanish (1,343,225) and English (1,073,209). The following are the main features of the corpus:

Documents belong to different institutions in the United Nations, namely: The Security Council, the Economic and Social Council, the General Assembly, UNESCO, etc.

Given the nature of the documents, a high frequency of Named Entities is observed, especially, proper nouns (person names, institutions and toponyms), acronyms and date expressions.

Texts represent a modern standard use of the language, although the language of the legal domain is predominant in many cases.

2. Processing and Alignment

The compiled documents from the previous stage are the input to the this stage, which consists of processing the corpus in order to provide an aligned version of the corpus. However, to achieve this output, the input passes through different modules of processing; mainly the basic processing module and the alignment module. Each is carried out on a monolingual basis.

2.1. Basic Processing

The basic processing includes three submodules:

1. Conversion from pdf to text format
2. Segmentation
3. Tokenization

2.1.1. Conversion from pdf to text format

In this module, the documents in pdf format are converted into plain text format and are saved as UNICODE. The latter option is justified if we take into account two reasons. First, it is an encoding scheme that allows the use of different writing systems. This is a critical feature in our case since we are dealing with three languages applying various writing systems. Second, UNICODE is the standard character set in XML.

Despite the simplicity of the conversion process, there are some of observations that should be highlighted. For the conversion of the Arabic documents, it is necessary to use a version of Adobe Acrobat Reader with special support for Semitic languages and bi-directional texts. Besides, during the conversion process, some combinations of characters are not recognized and thus it is necessary to replace them with their corresponding characters. For example, after conversion, the words "المجلس" (Council) or "نهاية" (End) appear as "اٍلس" and "ٍاية" respectively. In such cases, it is important to replace the unrecognized characters with the right ones. It has also been observed that many spaces have been replaced by double spaces in the conversion. Such observations are considered a source of noise that might affect the next stages of tokenization and segmentation.

2.1.2. Segmentation

The segmentation is carried out on the paragraph and the sentence levels. For the segmentation of the Spanish subcorpus, we used the tool available at the LLI-UAM, while for the English, we used the tool available in the Wraetlic[1] package. In case of the Arabic, we developed a simple rule-based segmentation tool considering the main features of the Arabic text. One of the main challenges, in this respect, is the bi-directional feature in the Arabic text, especially when a numeric expression precedes the end of the sentence. This difficulty can be explained if we take into consideration that numbers in Arabic are written from left-to-right while the alphabetical characters are written in the opposite direction. In cases like the example stated above, the numeric expression at the end of a sentence alters the position of the sentence making it necessary to adjust its position before segmentation.

On the other hand, the results of this process applied on a fragment of the corpus reveal that the relation between the sentences in the translation is not usually one-to-one. At the same time, it reflects some basic features of each language. For example, the lower number of sentences in Arabic indicates the tendency of the Arabic to merge sentences, while the English tends to use more sentences and the Spanish adopts an intermediate position.

2.1.3. Tokenization

For the tokenization, we used three language dependent tools. For Spanish, we applied the tokenizer available at LLI-UAM. For English, we used the Wraetlic tool. For the Arabic, we developed a simple tokenizer. Comparing the results of the tokenization process concerning the frequencies and the relation *token-type*, we found out that there are three main sources of textual noise in the three subcorpora:

- Noise due to altering the page format during conversion into text format. For example, the previously mentioned problem of double spaces or the documents' headers that, during conversion, passes to form part of the body text.
- Noise due to misspelling of some words
- Noise due to inherent features of the writing system of a certain language.

The first and second sources are language independent and, thus, cases were observed in the three subcorpora. Cases related to inherent features of the writing system, however, were obvious in the Arabic corpus. The absence of diacritics, on one hand and the use of *tatweel* (a character used sometimes in intermediate positions to lengthen the word), on the other hand, are basic sources of textual noise in the text.

The absence of diacritics decreases the number of types in its relation with tokens. For example, the type "تـقـدم" [noun: progress (تَقَـدُّم takaddom)/verb:to progress (takaddama تَقَدَّمَ)/to present (tokaddiتُقَدِّم)/to face (tokdim تُقْدِم)] without diacritics is highly ambiguous because it might be read in different ways. Therefore, what is considered in tokenization as one type, in fact, it represents various types.

The impact of missing diacritics might be compared to homographs in Spanish and English since they alter, in the same way, the frequencies of types. For example, words like "pack" in English could refer to two types "pack " as a noun or the verb "to pack". In Spanish, a word like "sobre" might refer to the preposition (on) or a

[1] This package developed by E.Alfonseca (UAM) is available in: http://www.ii.uam.es/~ealfon/eng/research/wraetlic.html

conjugated form of verb `“sobrar” (to give).` Despite of this similarity, in Arabic it is not a matter of homographs since they are different words with different spelling. It is the writing practices that drop the diacritics.

The *tatweel* has an opposite effect, since it increases the number of types when, in fact, it is only one type. For example, the word “المجلس” in the same corpus appears in different ways due to the use of *tatweel*. The following are some frequencies of the mentioned example:

المجلـــس	170	المجلــــــس	88
المجلس	124	المجلـــــــس	43

Handling the absence of diacritics in this stage is not feasible since it requires a pre-module of morphological analysis and disambiguation to assign the correct diacritics to each word. In this case, this would suppose a problem of circularity, as we need first to tokenize in order to continue with the tagging process. Consequently, in this stage, we decided to limit the results of tokenization to the simple task of preparing the text for further processing, without getting into other level of analysis. Nonetheless, the *tatweel* phenomena could be easily handled in this stage, so we decided to eliminate all the *tatweel* characters.

2.2. Alignment

The input of this module consists of the three subcorpra tokenized and segmented. The first output consists of three files each containing the alignment results of each pair of corpora, namely, the alignment of the Arabic-Spanish, Spanish-English and Arabic-English.

The alignment is carried out on the sentence level. After manual validation, the second output consists of a single XML file formatted according to the Translation Memory eXchange (TMX) standard.

The basic unit of the TMX standard is the translation unit <tu> corresponding, in this case, to the sentence. Each translation unit includes three sub-units representing the sentence in each of the three languages in concern.

To obtain this output, it is important to point out the following:

1. Alignment technique
2. Evaluation of the technique: results obtained in the three language pairs

2.2.1. Alignment Technique

The alignment technique is based mainly on the statistical model of Church and Gale (1991; 1993) where the correlation between the sentences’ length is the basic factor in the alignment. However, the results obtained by applying the statistical model showed high rate of errors, so we decided to include lexical information as anchor points. Given the high frequency of Named Entities in the corpus, they were tagged in the different corpora to be used as anchor points.

The tagging scheme adopted in the annotation of the Named Entities differentiates between seven basic categories: Person Name, Job, Toponym, Acronym, Institution, Event and Date. The category is indicated in an attribute called type. Each Named Entity is given a unique ID in the three languages. The following tag represents this scheme:

```
<ne type= “”id = “” ></ne>
```

Using Named Entities as anchor points is a novel approach as NE is a concept widely used in the field of Information Extraction, but not in alignment (Samy *et al.* 2005). Compared to the cognate technique frequently applied in the alignment studies, Named Entities tagging proved to be a more comprehensive approach. First, it is considered an added value to the multilingual corpus. Second, the Named Entity is a broader concept than cognates and, thus, offers a wider coverage. Finally, it is applicable to languages which do not share any phonetic or orthographic similarities.

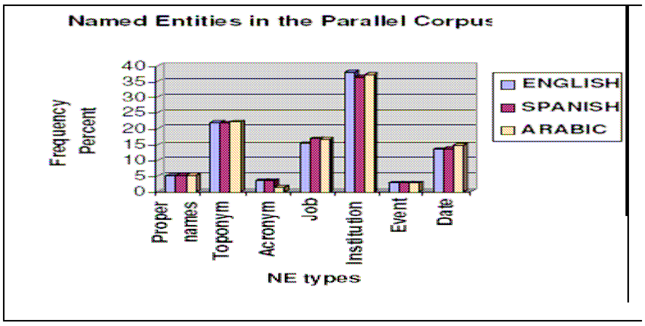

Figure 1. NE Distribution in the multilingual corpus

Once the NE are tagged, each pair of the subocpora passes through the alignment module which makes use of the statistical information together with the lexical information provided by the anchor points.

2.2.2. Evaluation of the technique: results obtained in the three language pairs

Results obtained by this method were satisfactory aligning more than 90% of the corpus, although the percentages differed from one language pair to another. These results were evaluated against a gold standard consisting of a manually aligned fragment (1200 pairs of sentences). The following graphs show the results of the automatic alignment compared to the manual alignment in the test data.

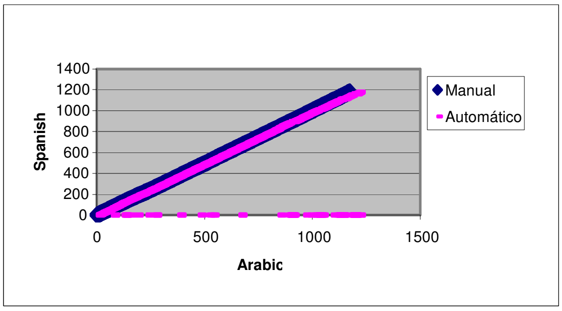

Figure 2. Alignment results (Arabic-Spanish)

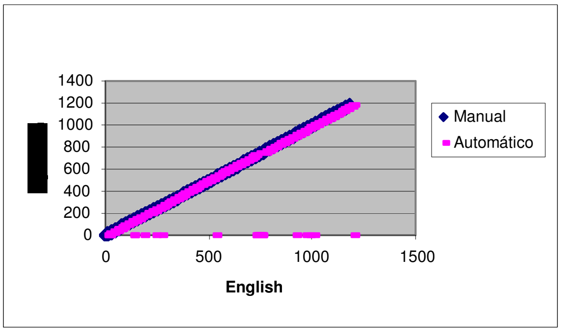

Figure 3. Alignment results (English-Spanish)

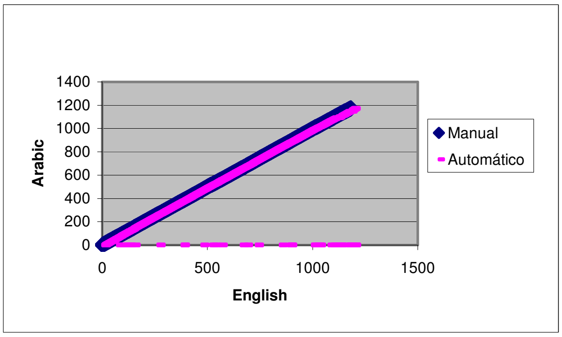

Figure 4. Alignment results (English-Arabic)

Evaluating the alignments in the three language pairs shows that the automatic alignment achieved the best results in case of English-Spanish with a percentage of 97,8% followed by the English-Arabic pair 95,4% then by the Arabic-Spanish 92%.

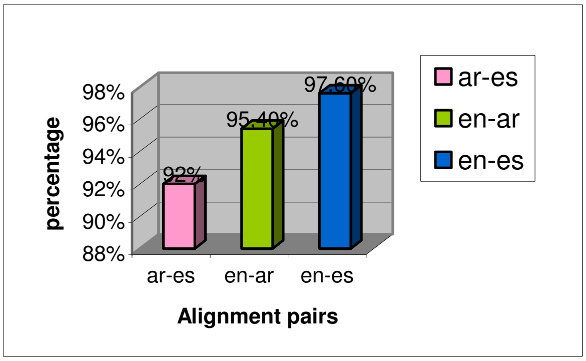

Figure 5. Alignment results in the different language pairs

Analyzing these results arises a number of questions related to the factors affecting the alignment process. In this paper, we would like to highlight two basic assumptions.

The first one has to do with the similarity between the languages and how it affects the alignment process. In other words, is there a correlation between the similarity between the languages and the accuracy of the alignment?

The second question tackles the issue of the direction of the translation. Are direct translations aligned better than indirect ones?

Regarding the first question, the results obtained in this stage show that the best results were obtained among the most similar language pair; English-Spanish. Despite these results, we cannot assure this fact, as it needs to be tested on another language pairs.

Concerning the factor of the nature of the translation with respect to its direction, the percentages show that the highest error rate was reported in the Arabic-Spanish pair. This proves this assumption to be true. Besides, in the UN document, it is explicitly mentioned that both the Spanish and the Arabic versions were translated from the original English document. One more evidence on this fact is that the percentage achieved in the Arabic-Spanish 92% is the result of the difference between the English-Spanish pair 97,8%and the English-Arabic pair 95,4%.

In spite of the stated hypothesis, we, again, insist that these are preliminary findings and further research is needed to prove these assumptions.

Another observation that we consider highly relevant is the multiple alignments vs. the single alignments. By multiple alignments, we mean the alignments where the relation between the sentences is one-to-two or two-to-one. Sometimes cases of many-to-many alignments could be reported. On the other hand, single alignment is used to refer to cases of one-to-one alignments.

The data analysis shows that the frequency of multiple alignments and the error rate are highly correlated. In that way, the more the multiple alignments, the more the errors. However, the technique applied was able in many cases to detect multiple alignments. For example in the English-Spanish pair, the technique applied correctly aligned 75% of the (2-1) cases and 35% of the (1-2) cases. Furthermore, in the English-Arabic pair, the percentages were 34% and 36% respectively and in the Arabic-Spanish pair the percentages were 50% and 35%.

The final output after validating the results is an TMX that structured as shown in the following fragment

```
<tu tuid="205" datatype="Text">
   <tuv xml:lang="ar">
    <seg>
   وزار رئيسا إندونيسيا و البرتغال و
رئيس وزراء أيرلندا الإقليم
خلال تلك الفترة.
    </seg>
   </tuv>
   <tuv xml:lang="es">
     <seg> Los Presidentes de Indonesia y
Portugal y el Primer Ministro de Irlanda
visitaron el Territorio en ese período.
     </seg>
   </tuv>
<tuv xml:lang="en">
     <seg> The Presidents of Indonesia and
Portugal and the Prime Minister of Ireland
visited the Territory during that period.
    </seg>
   </tuv>
</tu>
  <tu tuid="206" datatype=”Text”>
   <tuv xml:lang="ar">
    <seg>وأشير
              إلى أن الحالة الاقتصادية والاجتماعية
                                          كانت
              موضع قلق المتكلمين في تلك الجلسة.
    </seg>
   </tuv>
   <tuv xml:lang="es">
    <seg> En la sesión se señaló que
causaba preocupación la situación económica
y social imperante.
    </seg>
  </tuv>
 <tuv xml:lang="en">
    <seg> The economic and social situation
was mentioned as an area of concern by the
speakers at the meeting.
    </seg>
  </tuv>
</tu>
```

3. POS Tagging

In the previous stage we managed to provide an version of the corpus aligned on the sentence level. In this stage, our main goal is to tag the three subcorpora on the POS level. The output consists of the three subcopora tagged with POS in XML format.

Given the nature of this task and its dependency on the language, the tagging is done on monolingual basis. Three taggers are used. For Spanish tagging, we used the POS tagger developed in the LLI-UAM and based on the morphological analyzer GRAMPAL (Moreno Sandoval *et al.* 2005). For English tagging, we used the tool provided by the Wraetlic tools and finally for Arabic, we developed a rule-based tagger.

Since we are dealing with three different languages; a romance language, an anglo-germanic and a semitic one, each represents a number of features on the morphosyntactic level requiring tools that could efficiently handle these features.

Starting with the Arabic, in the design of the tagger we respected the principles of the grammatical tradition (Khoja 2001). According to these principles, Arabic distinguishes between three main wordclasses; noun, verb and particle.

The following table presents the distribution of the different categories in the Arabic corpus.

Category	Percentage
Nouns without clitics	15,5
Verbs	7,6
NE Tokens	6,1
Punctuation	9,2
Closed Categories	17,3
Noun+enclitics	2,5
Proclitics+Nouns	18,5
Proclitics+ Nouns+Enclitics	0,5
Proclitics+Closed Categories	2,6
Closed Categories+Enclitics	1,3
Proclitics+ Closed Categories+ Enclitics	0,2

Table 1. Distribution of POS categories in the Arabic corpus

The results shown in the table reflect the high use of clitics in the Arabic language. This phenomenon reveals the complexity of Arabic word structure, which in some cases might be made of up till 4 word classes. Considering this feature the classification is carried out with especial emphasis on the use of clitics. However, a detailed discussion of the POS tagger of the Arabic is out of the scope of this paper. Thus, we limit our discussion to the results. The following is a fragment of the Arabic corpus after POS tagging.

```
<p id="1">
<s id="1">
<tok type="comp">
<orth>وأشار </orth>
<desc>conj+verb</desc>
<part type="conj"> و </part>
<v  lema="أشار"  raíz="شير"  temp="pasado"
per= "3" num="sg" gen="m">أشار </v>
</tok>
<ne type= "job" id="61">وكيل الأمين العام
</ne>
<punct>،</punct>
<tok type="comp">
<orth>بوجه</orth>
 <desc>prep + noun</desc>
 <part type="prep">ب</part>
 <noun  type="common"  gen="m"  num="sg"
lema="وجه" raiz="وجه">وجه</noun>
</tok>
 <tok type="senc">
<orth>خاص</orth>
 <noun type="adj" gen="m" num="sg" lema="
خاص/خاصة" raiz="خص">خاص</noun>
</tok>
<punct>،</punct>
<tok type="senc">
<orth>إلى</orth>
 <part type="prep" lema="إلى">إلى</part>
</tok>
<tok type="comp">
<orth>التقدم</orth>
 <desc>art + noun</desc>
 <part type="art">ال</part>
```

Regarding the Spanish POS tagging, tags are classified into 17 categories. The following table represents the distribution of these categories in the corpus

Category	Percentage
PREP (preposition)	16,5
N (noun)	15,7
ART (article)	15,6
NE (named entity)	8,7
V (verb)	8,4
PUNCT (punctuation)	8,2
ADJ (adjective)	4,7
C (conjunction)	3,8
Q (quantifier)	2,3
P (pronoun)	1,8
NUM (number)	1,01
ADV (verb)	0,96
AUX (auxiliary)	0,89
POSS (possessive pronoun)	0,86
REL (relative pronoun)	0,71
DEM (demonstrative pronoun)	0,34
MD (discourse marker)	0,17

Table 2. Distribution of POS categories in the Spanish corpus

The fragment included below is part of the Spanish corpus after tagging.

```
<p id="1">
<s id="1" nt="33">
   <w   cat="ART"   lem="el"   gen="masc"
num="sing"> El </w>
   <ne  type="job"  id="61">  Secretario
General Adjunto </ne>
   <w cat="P" lem="se"> se </w>
   <w        cat="V"        lem="referir"
tie="indf_ind"     num="sing"     per="3">
refirió </w>
   <w cat="PREP" lem="en"> en </w>
   <w cat="N" lem="particular" gen="masc"
num="sing"> particular </w>
   <w cat="PREP" lem="a"> a </w>
   <w   cat="ART"   lem="el"   gen="masc"
num="plu"> los </w>
   <w  cat="N"  lem="progreso"  gen="masc"
num="plu"> progresos </w>
   <w  cat="ADJ"  lem="logrado"  gen="masc"
num="plu"> logrados </w>
   <w  cat="PREP"  lem="en  relación  con">
en relación con </w>
   <w    cat="ART"    lem="el"    gen="fem"
num="sing"> la </w>
   <w  cat="N"  lem="iniciativa"  gen="fem"
num="sing"> iniciativa </w>
   <w cat="PREP" lem="de"> de </w>
   <w    cat="N"    lem="paz"    gen="fem"
num="sing"> paz </w>
   <w cat="PREP" lem="de"> de </w>
   <ne type="top" id="49"> Djibouti </ne>
```

Finally, the English POS tagging done by the Wraetlic tools provides the following output.

```
   <w c="w" pos="PRP">he</w>
     <w c="w" pos="RB">also</w>
     <w c="w" pos="VBD">indicated</w>
     <w c="w" pos="IN">that</w>
     <w c="w" pos="DT">the</w>
     <w c="w" pos="NN">initiative</w>
     <w c="w" pos="VBD">had</w>
     <w c="w" pos="VBN">been</w>
     <w c="w" pos="RB">well</w>
     <w c="w" pos="VBN">received</w>
     <w c="w" pos="IN">by</w>
     <w c="w" pos="JJ">Somali</w>
     <w c="w" pos="NN">society</w>
     </s>
```

4. Conclusions and Future word

In this paper, we introduced part of the results of an on-going research in the LLI-UAM to build a multilingual parallel corpus. A fragment of the corpus is freely available online through the webpage of the LLI-UAM http://www.lllf.uam.es/

The preliminary results reported in this paper helped us in putting hands-on key issues in alignment of multilingual corpora, in general, and Arabic language processing, in particular.

The nature of the corpus and the diversity of the languages gave us the opportunity to address a number of problems and challenges from both the technical and the linguistic point of view. However, the results reached in this stage are mainly descriptive and, thus, further research is required to test the different assumptions and observations, we have been pointing out all through this paper.

Future studies will focus on extending the corpus. The extension is considered in a vertical and a horizontal way. The horizontal extension deals mainly with expanding the size of the corpus, on one hand and including more languages, on the other hand.

Regarding the vertical extension, from a technical point of view, we would consider other levels of processing such as basic chunking and semantic annotation. Also, testing techniques of lexicon extraction and multi-word expressions across the different languages are subjects of interest in the future.

From the linguistic point of view, this corpus is a valuable resource for the translation and the contrastive studies. At the same time, we would study the role of this kind of resources in second language teaching and acquisition.

5. References

GALE, W.A. & K.W. Church (1991): A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of 29th Annual Meeting of the Association of Computational Linguistics ACL' 91*, pp.177-184.

GALE, W.A. & K.W. Church (1993): A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* (19), pp. 75-102.

GARSIDE, R., J. Hutchinson, G. Leech, A.M. McEnery & M. Oakes (1994): The exploitation of parallel corpora in projects ET10/63 and CRATER. In Jones D.B. (Ed.), *New Methods in Language Processing*, Manchester: UMIST, pp. 108-115.

IDE, N. & J.Véronis (1994): MULTEXT: Multilingual Text Tools and Corpora. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, Kyoto, Japan.

KHOJA, S. (2001): APT: Arabic part-of-speech tagger. In *Proceedings of Student Workshop (NAACL2001)*, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States.

KOEHN, P. (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit MT Summit X- 2005*, Phuket Islan, Thailand.

MORENO-SANDOVAL, A., G. De la Madrid, M. Alcántara, A. González, J.M. Guirao & R. De la Torre (2005): The Spanish Corpus. In CRESTI, E. y M. Moneglia (Eds.), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam: John Benjamins Publishing Company, pp. 135-161.

SAMY, D., A. Moreno-Sandoval & J.M. Guirao (2005): A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus (Spanish-Arabic). In *Proceedings of International Conference on Recent Advances on Natural Language Processing RANLP 2005*, Borovets, Bulgaria, pp. 459-465.

SAMY, D. (2005): Recursos bilingües de ingeniería lingüística para el procesamiento de español y árabe. PhD Thesis. Autónoma University Madrid.

TIEDEMANN, J. & L. Nygaard (2004): The Opus corpus- parallel and free. In *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC'04*, Lisbon, Portugal.