Construction of a Bilingual Arabic-Spanish Lexicon of Verbs Based on a Parallel Corpus

Doaa Samy[†] **Antonio Moreno-Sandoval**[†] **José M. Guirao**[‡]

†Laboratorio de Lingüística Informática
Universidad Autónoma Madrid
Cantoblanco 28049 Madrid, Spain
{doaa , sandoval}@maria.lllf.uam.es

‡Dept. of Software Engineering
University of Granada
guirao@ugr.es

Abstract

Parallel corpora are considered an important resource for the development of linguistic tools. In this paper our main goal is the development of a bilingual lexicon of verbs. The construction of this lexicon is possible using two main resources: I) a parallel corpus (through the alignment); II) the linguistic tools developed for Spanish (which serve as a starting point for developing tools for Arabic language). At the end, aligned equivalent verbs are detected automatically from a parallel corpus Spanish-Arabic. To achieve this goal, we had to pass through different preparatory stages concerning the assessment of the parallel corpus, the monolingual tokenization of each corpus, a preliminary sentence alignment and finally applying the model of automatic extraction of equivalent verbs. Our method is hybrid, since it combines both statistical and linguistic approaches.

1. Arabic and Corpora

In this introductory section, we would like to highlight the state of art in the field of Arabic corpora. The actual linguistic panorama reveals an increasing interest for building Arabic corpora.

Considering the Arabic available corpora, there are three main written sources:

1. The Arabic Newswire, built by the LDC at Pennsylvania University. It is a compilation of articles from Agence France Presse and it consists of 76 million words.
2. Articles from the Lebanese newspaper *Al-Nahar*, with 140 million words. Available through ELRA.
3. Articles published in Al-Hayat newspaper, compilated by De Roeck and Goweder (2001). Also available through ELRA.

Concerning the spoken corpora, the LDC has compiled two phone-recording corpora of Egyptian spoken Arabic (CALLHOME and CALLFRIEND).

On the other hand, the parallel corpora are of especial importance for the multilingual language processing tools. The survey of the state of art in this aspect does not show any evidence of studies concerning the Arabic language in parallel corpora. The only evidence in this aspect is the work of Resnik and Smith (2003) for the STRAND project concerning the retrieval of parallel corpora from Internet. In the case of the Arabic language, the system was able to locate 2,190 URL pairs for English-Arabic documents.

2. The Spanish-Arabic parallel corpus

The above survey shows the absence of the Arabic language from the panorama of cross-lingual parallel corpora. This can be explained if we take into consideration the following facts:

- Most of the computational and corpus- linguistic studies concerned with Arabic have studied this language in comparison mainly with English, but also with French.
- Spanish, on the other hand, has been studied mainly in comparison with English, and with other European languages.

2.1. Building the corpus

In this section, we will briefly discuss the central features of the corpus and the selection criteria.

In the compilation phase our main objective was to build a parallel corpus, that is, "a set of L1 texts and an equivalent set of L2 translations of L1" (McEnery, 1997). In other words, "a text which is available in two (or more) languages" (Somers, 2001).

The first task consisted in locating documents in Spanish and Arabic through the World Wide Web. The results of the first search was not satisfactory considering the quantity and the quality. The available texts in Arabic with its translations in Spanish and viceversa are relatively scarce. Besides, the quality of the translation either Spanish-Arabic or Arabic-Spanish was not appropriate to allow a linguistic study. In a second round, search results were much better since it met both criteria quantity and quality. A set of official texts of the United Nations were located and compiled, since both Spanish and Arabic are, among others, UN official languages,.

2.2. Corpus characteristics

Through the available UN documents, it was possible to build up a parallel Spanish-Arabic corpus consisting mainly of annual reports of different UN institutions, such as the Security Council. All texts are equivalent in both languages, with a total size of about 2 million tokens. The corpus reveals the following features:

1. It is a representation of modern standard Arabic and Spanish, used in formal official documents.
2. As UN documents, the quality of translation is guaranteed.

3. The abundant use of Named Entities (NE): proper names, dates, countries, etc.

2.3. Corpus Assessment

2.3.1 Monolingual tokenization and segmentation

Tokenization is the first step in the process. The goal is to identify the *linguistic tokens* from the *graphic tokens.* A tokenizer for each language has been developed, reflecting orthographic and textual idiosyncrasies of written Spanish and Arabic. “The tokenization process depends strongly on the type of text” (Grefenstette 1999: 118) and even more depends on languages from different linguistic and cultural traditions. Regardless of the language, the output of the tokenization should include the following:

1. Identification of token boundaries, abbreviations, and punctuation marks.
2. Recognition of numbers, sometimes written in different alphabets.

Segmentation, on the other hand, is the process of identification of structural linguistic units, namely, paragraphs and sentences. This process is based on the previous recognition of linguistic tokens, and is the basis for the parallel text alignment.

Arabic tokenization and segmentation

Tokenization: The original documents were in pdf format, so the first step consisted in converting the documents from pdf format to text format. The conversion was possible using a special version of the Acrobat Reader; Acrobat Reader Middle East version, since it provides the appropriate support for Arabic text fonts and for bi-directional texts. All the texts were saved as Unicode texts to avoid problems of character encoding, as the parallel corpus uses two different writing systems.

The Arabic tokenizer was developed using PERL with Unicode support. The tokenization process had to take into account a series of features in the Arabic text:

1- During the conversion process from pdf format to text format, many single spaces where substituted by double spaces. Such a feature may be a source of noise when detecting the word boundaries. To avoid this problem, all double spaces were substituted by single spaces.

2- The *tatweel* is a common phenomenon in Arabic texts. It consists of the character “ـ” which is used for esthetical purposes. For example, the preposition “من” (*of*) may appear in different forms depending on the use of the *tatweel*, resulting in various tokens for the same type. For example, in a sample of the Arabic corpus(256,000 words):
“من” occurs with a frequency of 1957, while
“مـن” occurs 1921 times and
“مــن” occurs 961 times.
In the three cases it is the same word. To avoid this problem, we eliminated all cases of *tatweel.* This was reflected clearly in the total number of types after and before eliminating the *tatweel.* In the sample, the number of types before eliminating the *tatweel* was 37,161. After eliminating the *tatweel*, this number was reduced to 18,949.

3- The Arabic conjunction “و” (*and*) in almost all cases appears combined to the following word without a separating space. This feature results in problems of word boundary identification and ambiguity in the posterior word class tagging. However, eliminating or separating the conjunction may not be an appropriate solution at this stage.

4- The punctuation marks are eliminated except the full stops, since they are crucial for the next stage of segmentation and the commas as they are used as indicators of verbs in the following verb tagging module.

Segmentation:. We would like to highlight two main characteristics of the Arabic text:

1- Abbreviations are so rare in the Arabic texts in general. In our sample no cases of abbreviations were detected. That is why the use of full stops is exclusive for indicating sentence boundaries.

2- The use of numbers within the Arabic text caused much noise in the detection of the sentence boundaries. The Arabic text follows a right-to-left direction, while the numbers are written in a left-to-right direction. If the sentence boundary coincides with a number (either [٠١٢٣٤٥٦٧٨٩] or the western numbers [0-9]), the direction changes causing alterations in the position of the full stops, ending up as a source of noise during the segmentation process.

The experiment reported in this paper was carried out in a sample of the corpus. The results of tokenization and segmentation are summarized in the following table:

Language	Spanish	Arabic
No. of Tokens	39,496	25,144
No. of Sentences	1,168	1,007
No. of Paragraphs	709	528
Average tokens/sentence	33.64	24.96
Average tokens/paragraph	55.71	47.62

Table 1: Tokenization and Segmentation Results

The results in the table reveal a noticeable difference between the number of tokens in the Spanish corpus and those in the corresponding Arabic corpus. This is an expected phenomenon, due to the nature of each language. The Arabic language tends to combine more often different morphemes in a single token, such as articles, object pronouns and prepositions.

2.3.2. Partial tagging

After the tokenization and segmentation, a partial monolingual tagging is necessary for the parallel text alignment, since a previous word anchorage is needed.

Before discussing the tagging procedure, we will point out some basic facts about the Arabic language and how these facts affected our experiment.

- The Arabic language as a Semitic language is rather different from Spanish language. These differences are observed at the structural and grammatical levels. In that way, our attempt for the alignment and tagging of a parallel corpus Spanish-Arabic using the Spanish tools is considered the first attempt to study this linguistic pair from a computational point of view.

- The Arabic language does not differentiate between an upper case and a lower case.

- Arabic is written from right to left.

- Arabic is a highly inflected language, where the verb roots and patterns are the basis for the different morphological categories. That is why building a lexicon of verbs would be a very valuable resource for Arabic Language Processing.

- Similar studies concerning the Semitic languages (e.g. Hebrew) point to the fact that no statistical procedures, especially the alignment is possible without some normalizing pre-processing; lemmatization[1] (Choueka et al., 2000). However, the principal motivation behind this experiment is based on the idea of how to use parallel corpora and previously developed tools for a language L1 (Spanish) as resources for developing NLP tools for another language L2 (Arabic). In this way, the tagged Spanish corpus served as a starting point for developing our Arabic tagger.

The Spanish corpus was morphologically analyzed and lemmatized using previously developed tools such as GRAMPAL. At this stage, the partial tagging of the Spanish corpus was mainly concerned with the NE, since they provide the basis for the alignment process. Identifying the NE candidates in the Spanish text was mainly based on the orthographic aspects; the use of the upper case and the common dates patterns. The NE include the following categories:

- Country Names and Toponyms
- Proper Names and dates

In the case of the Arabic corpus, since there are no tools available, we developed an automatic tagger for the annotation of the following categories:

- NE
- Closed categories (prepositions, pronouns and conjunctions)

Named Entities: The development of the Arabic tagger for NE was based on the results of the Spanish tagging. Once the NE are identified in the Spanish Corpus, a list of these NE is generated and the corresponding Arabic NE are provided by a linguist. When provided, the Arabic tagger makes use of this list to annotate the NE in the Arabic text. This procedure was adopted since the Arabic language does not differentiate between lower case and upper case. Thus, a formal recognition of NE in the Arabic text would be impossible.

3. Sentence Alignment

The alignment achieved in this experiment was done on the sentence level. We developed an alignment tool which makes use of the high frequency of previously annotated equivalent NE in both texts.

Previous studies have proven the utility of establishing anchor points in order to reduce the noise during the alignment process. In our case, we made use of the paragraph subtitles where the subtitles are NE (Country names and Toponyms). At this stage, we identified 40 anchor points in the parallel test corpus.

Based on these anchor points, in the following steps, we try to find alignments which maximize the correspondence between the NE, using dynamic programming techniques as used by Wu & Xia (1995).

At first, we established a one-to-one sentence alignment. At this stage, the model was able to detect 681 alignments. In a posterior stage we implemented other techniques to allow a multiple sentence alignment. The results of the multiple alignments was 28, achieving in this way a total of 709 alignments.

4. Extraction of equivalent verbs

4.1. Monolingual annotation

Spanish verb tagging

Our starting point is the lemmatization of the Spanish verbs in the corpus. For this task we use GRAMPAL (Moreno 1991; Moreno and Goñi 1995), a morphological processor based on a rich morpheme lexicon of over 50.000 lexical units, and morphological rules. The original system has been extended to POS disambiguation and to annotate spoken Spanish (Moreno and Guirao 2003).

GRAMPAL is applied on the NE annotated corpus providing, as output, the Spanish corpus with the verbs tagged and lemmatized.

Arabic verb tagging

In the case of the Arabic corpus, since no tools are available for verb recognition, we developed a programme for identifying the verbs in the Arabic text. The algorithm consists of the following steps:

1- Named Entities are annotated in the Arabic corpus
2- Closed categories including prepositions, pronouns and conjunctions are annotated.
3- The remaining untagged words are passed to a list.
4- A set of grammatical and contextual rules for detecting Arabic verbs are implemented in such a way that the elements in the list of untagged words are tested against these rules and they are given different weights indicating their probability to be verbs.
5- The grammatical rules consist of verb affixes which are implemented through regular expressions. But since the Arabic language is highly ambiguous. These rules were combined with a set of contextual rules, such as:

a) Verbs never start with an article
b) Relative pronouns are almost followed by verbs
c) Verbs often appear in the first position of a sentence (after the full stop), or a phrase (after the comma).

6- The words with higher weights are selected as candidate verbs. A linguist verifies the candidate list manually and the final candidates are then passed to the tagger, which identifies their occurrence in the corpus.

The verb tagger was able to identify a total of 1804 verbs in the sample corpus with an average of 1.79 verb/sentence. In the Spanish corpus, the total number of verbs detected is 1168, with an average of 1verb/sentence.

Evaluation: In order to evaluate the Arabic verb detection, we created a sample golden standard of Arabic verbs. The procedure was carried out as follows:

1- We took approximately one-third of the corpus as a sample, and a linguist manually extracted all verbs in that test corpus. The result was a list of verbs per sentence. In total, 302 sentences (from sentence 1 to

[1] Those authors defend the strategy for Hebrew, but the claim can be applied to Arabic and Spanish, in a natural way.

sentence 100; from 300 to 400; and from 800-900) were verified manually, and 676 verbs were extracted (that is, the *total number of verbs* in the sampling). The verb/sentence ratio in our test corpus is 1.79.

2- Automatically every verb in the golden list was searched in the actual annotated corpus. Every time a search matches, the counter of *correct detected* is augmented by one.

3- Previously, we counted the number of verbs detected by our Arabic verb tagger in the test corpus (559 verbs): the *total number of verb candidates*.

Calculating both, the recall and the precision we obtained the f-measure. The following table shows the evaluation results of the verb detection process.

Recall	Precision	F-measure
75.74%	91.59%	82.91

Table 2: Recall and Precision

3.2. Bilingual extraction of Equivalent Verbs

We use a standard Mutual Information measure for extracting verb pairs as candidates to be mutual translation of each other. The pair consists of previously tagged verbs in the Spanish corpus, and candidates to verbs selected and tagged in the Arabic corpus.

Mutual Information is defined by

$$Mi(v_s, v_a) = \log_2 \frac{prob(v_s, v_a)}{prob(v_s)\,prob(v_a)}$$

To avoid unreliable matches when the counts are small, a t-score is used to filter out insignificant mutual information values. When t-score is less than 1.65 corresponding to a confidence level of 95% (Fung 1995), the filter is applied.

At the end, our program produces a list of suggested Arabic words for each Spanish verb in the test corpus. The following table shows a sample of the results we obtained:

Verb Pair	Co-occurrence	Mutual Information	*t*-score
أعرب<->expresar	(45) : (6)	6.20	2.41
دارت<->celebrar	26) (1: (4)	5.89	1.96
اســتمع<->escuchar	(26) : (3)	5.72	1.69
قدمها<-> presentar	27) : (3)	7.01	1.71
نــاقش<-> analizar	(8) : (3)	7.64	1.72
عقد<->celebrar	(126) : (4)	5.12	1.94

Table 3: Sample of Equivalent Verbs

4. Conclusions and future work

Bilingual lexicon extraction has been used for different purposes: to find terminology units and their translations (Gaussier et al 2000), to augment dictionaries for Cross-Language Information Retrieval (Brown et al. 2000); or most frequently, to help human lexicographers. In our case, we want to apply the technique in building a morphological processor for Arabic (a lemmatizer and a POS tagger) based on a rich lexicon. Since we already got tools for Spanish (a lemmatizer, an NE recognizer), we used them as starting point for the alignment and lexicon extraction. In the experiment, we have shown that a few rules combined with manual selection can provide sufficient candidates for being annotated as Arabic verbs. With help of the Spanish verbs in the parallel corpus, the extraction algorithm identifies the equivalent Arabic ones. As a result, we obtain a list of verb forms, and their equivalent in the other language.

We will repeat this method to the complete parallel corpus, extending to the rest of inflectional categories (nouns and adjectives). In an incremental manner, the lexicon-based morphological tagger will be augmented. POS and sense disambiguation will be treated in a subsequent phase.

5. Acknowledgements

This work was funded by a scholarship from the Spanish Agency of International Cooperation AECI.

6. References

Brown, R., J. Carbonell, and Y. Yang 2000. Automatic dictionary extraction for cross-language information retrieval. In Véronis (ed.) (2000)

Chouka, Y., E. Conley, and I. Dagan 2000. A comprehensible bilingual word alignment system: Application to disparate languages, Hebrew and English. In Véronis (ed.) (2000).

De Roeck, A. and Goweder A. 2001. "Assessment of a Significant Arabic Corpus". In *Workshop Proceedings of Arabic Language Processing: Status and Prospects,* Toulouse, France, pp:73-79.

Fung, P. 1995. A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. In *Proceeding of the ACL* (pp. 236-243)

Fung, P. 2000. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In Véronis (ed.) (2000).

Gaussier, E., D. Hull, and S. Aït-Mokhtar 2000. Term alignment in use. In Véronis (ed.) (2000).

Grefenstette, G. 1999. Tokenization. In van Halteren (ed.) *Syntactic Wordclass Tagging*. Dordrecht, Kluwer.

McEnery, T. 1997. "Multilingual Corpora-Current Practice and Future Trends". In *13th ASLLB Machine Translation Conference*, London, pp. 75-86

Moreno, A. 1991. *Un modelo computacional basado en la unificación para el análisis y generación de la morfología del español*. Ph. D. Thesis. Madrid, Universidad Autónoma de Madrid.

Moreno, A. and J.M. Goñi 1995. GRAMPAL: a morphological model and processor of Spanish implemented in Prolog. In *Proceedings of GULP-PRODE*. Marina di Vietri, Italia.

Moreno, A. and J.M. Guirao 2003. Tagging a spontaneous speech corpus of Spanish. In *Proceeding of RANLP-2003*, Borovets, Bulgaria.

Resnik, P.and Smith, N. 2003. The Web as a Parallel Corpus*, Computational Linguistics* 29(3).

Somers, H. 2001. "Bilingual Parallel Corpora and Language Engineering". In *Anglo Indian Workshop "Language Engineering for South Asian Languages*" *LESAL*, Mumbai, April 2001.

Véronis, J. (ed.) 2000. *Parallel Text Processing*. Dordrecht, Kluwer.

Wu, D. and X. Xia 1995. Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon. *Machine Translation* 9 (3-4), pp 285-313.