Named Entities: Structure and Translation*

A study based on a Parallel Corpus (Arabic-English-Spanish)

Doaa Samy
Computational Linguistics Laboratory
Department of Linguistics
Autónoma University of Madrid
doaa@maria.lllf.uam.es

Introduction

The term Named Entity (NE), first introduced in 1995 by the Message Understanding Conference (MUC-6), is widely used in the field of Natural Language Processing and Information Extraction. Since 1995, studies concerning Named Entities have been gaining an increasing attention, as the introduction of NE modules (NE Tagging, NE Recognition or NE Extraction) has proved to be an efficient factor in enhancing IE systems (Sekine, 2004), (Grishman and Sundheim, 1996), (Hasegawa et al., 2004) and in improving alignment techniques (Melamed, 2001), (Samy et al., 2004).

This paper presents a study of the Named Entities based on a parallel aligned multilingual corpus (Arabic-Spanish-English). The study focuses on two main aspects; first, the way the different classes of Named Entities adopt different language-dependent patterns and, second, the strategies adopted for NE translation across the three languages (English-Arabic-Spanish). However, and as a starting point, in the first section, we will briefly discuss the state-of-art in Named Entities and the recent trends in translation studies. In the second section, we will explain the methodology adopted for the research, which depends mainly on the use of parallel corpus and previously developed tools for processing the Spanish and English text. The third section will focus on the characteristics of our multilingual parallel corpus and the NE tagging process for the three languages in concern. The analysis of translation patterns and strategies will be discussed in the fourth section, followed by the final section dedicated to the conclusion and future work.

1. State-of-art

1.1. Named Entities

Sekine (2004) defines Named Entities extraction from unstructured text as the task of recognizing information units, mainly names including person, organization and location names, and numeric expressions including time, date, money and percent expressions. This basic classification adopted during the nineties distinguished seven categories, grouped in two main classes; names and numeric expressions. Meanwhile, recent studies on Named Entities have proposed an extended NE hierarchy covering up to 200 categories (Sekine, 2004). Furthermore, other approaches implemented a NameNet model as a structured resource of name classes semantically related, following the WordNet paradigm (Morarescu and Harabagiu, 2004). However and in terms of

* This research has been supported by the grant TIN2004-07588-C03-02 (Spanish Ministry of Education and Science) and by an individual grant from the Spanish Agency for International Cooperation

feasibility, implementing such extended hierarchies is a complicated, expensive and time-consuming task, as it requires much annotation effort. Besides, it supposes a difficulty in deciding the right category for each Named Entity.

Dealing with Named Entities in the three languages English, Arabic and Spanish is a novel approach, since almost all the previous literature has tackled the issue of Named Entities from a monolingual point of view. Furthermore, the technical perspective has prevailed in NE studies, as they focused on developing different techniques for NE tagging, recognition and extraction. Thus, this paper suggests a new dimension for the study of Named Entities based on a translation perspective. Besides, it offers a multilingual approach dealing with three languages completely different in terms of typology, morphology, etc. Although English and Spanish might share some characteristics as Indo-European languages, their morphological and syntactic structures are quite different since English is an Anglo-Saxon language, while Spanish belongs to the Romance languages group. On the other hand, Arabic is a Semitic language representing a completely different set of features, not only on the structural or morphological levels, but even on the orthographic level, as it uses a different writing system. Dealing with the three languages supposed a series of challenges starting from the basic level of the writing systems passing through other levels of processing such as tokenization, segmentation and NE Tagging. In this paper, we will focus on the NE Tagging, as issues concerning tokenization and segmentation are beyond the scope of this study.

1.2. Translation studies and corpora

For a long time, studies concerning the translation have focused on long discussions about translation theories. However, recently there has been a growing interest in the field of translation studies to develop new approaches focusing on empirical data and data driven analysis based on real text. This interest has given rise to a recent trend known as corpus-based translation studies, which is gaining an increasing attention, since it "*reveals facts of the process and product of translation which are new, consistent, and based on solid empirical foundations*" (Laviosa, 2002). Furthermore, the corpus-based approach allows the possibility of analysing "*the individual particularities of specific pairings of languages in translation exchanges and the characteristics of translation as cultural interface at different times and places and under different cultural conditions*" (Laviosa, 2000).

The trend of corpus-based translation studies makes use of corpus linguistics as a methodology for addressing the translation issues and the role of the translator in this process from a philological and a linguistic point of view without taking into consideration a possible computational perspective which would highly benefit from the results conducted on the level of linguistic analysis.

2. Methodology

2.1. General Methodology: Bringing the Two Fields Together

In view of previous aspects concerning Named Entities and corpus-based translation studies, it is important to underline the fact that we are dealing with concepts belonging to different areas of research, which may be grouped in two main classes; the first

includes Information Extraction, Natural Language Processing and Computational Linguistics, while the second is concerned with the translation studies. Thus, our approach can be regarded as an intent to bring the two fields together.

Given this fact, we decided to adopt a new multi-dimensional methodology that aims at combining approaches from corpus-based translation studies, on one hand, and Information Extraction and Computational Linguistics, on the other hand. Regarding this fact, we would like to make clear that the concept we adopt for Parallel Corpora and Named Entities together with the methods applied for their classification and tagging are those applied in Computational Linguistics, while the concepts of linguistic patterns and translation strategies, belongs to the corpus-based translation studies. In our opinion, such a hybrid approach provides a mutual benefit. On one hand, translation studies would make use of tagging tools and IE systems, while at the same time such tools and systems would benefit from the results of the translation studies in implementing and improving their rules and grammars

However, such combination of approaches might result in some ambiguities regarding the use of terms and concepts. That is why it is indispensable to give clear and precise definitions of the basic concepts before proceeding on with our analysis.

Translation Corpus vs. Parallel Translation Corpus: Bilingual or multilingual "parallel translation corpus" are widely used concepts in the field of Information Extraction, Natural Language Processing and Computational Linguistics. On the other hand, the concept of monolingual "translation corpora" is used in corpus-based translation studies. Thus, the term "translation corpus" is ambiguous. In Computational Linguistics and NLP, a *parallel translation corpus* is a "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). This is the definition we adopt for the present study. However, in corpus-based translation studies a *translation corpus* is a monolingual corpus of translated texts since the main goal of these studies is to analyse the basic features of translated text either independently or in comparison with other non-translated texts of the same type resulting, in this way, in two types of corpora; translation corpus and comparable translation corpus.

2.2. NE Tagging Methodology: Parallel Corpus and Previously Developed Tools

From a technical point of view and with respect to the methodology applied for NE tagging in the parallel corpus, we decided to follow the most recent tendencies in Computational Linguistics and Natural Language Processing. Such tendencies aim at providing the necessary resources for languages with scarce or few ones in a fast and effective way by using parallel corpora and previously developed tools/resources for languages such as English or any other languages rich in resources. This is the case in the present study, since we used a parallel corpus and available tools for Spanish and English NE tagging to tag the Arabic Named Entities.

3. Tagging the Named Entities in The Parallel Corpus

3.1. The Parallel Corpus

A parallel multilingual corpus aligned on the sentence level and with the Named Entities tagged in each of the three languages was used for the present study. The corpus developed in our Laboratory consists of a collection of parallel texts in English, Arabic and Spanish obtained from the publicly available documents in the United Nations web page. The corpus is two million words' size, but the aligned part consists of almost 1200 pairs of sentences (English-Arabic), (English-Spanish) and (Arabic-Spanish). Despite the limited size of the corpus, the high frequency of Named Entities (average 2.4 NE/sentence) in the text and the quality of the translation guarantee a solid base for the conducted research.

The alignment was carried out using an enhanced version of the statistical model of Gale and Church. The version developed in the Computational Linguistic Laboratory combines the statistical information with the linguistic information provided by the tagged Named Entities considered as anchor points for improving the alignment process.

3.2. NE Tagging Scheme

Given the complication of the task of tagging the Named Entities in a multilingual corpus, we opted for a simple tagging scheme inspired in the guidelines provided by the basic Named Entity classification. We implemented a total of seven categories; six belonging to the name class and one belonging to numeric expressions. The six categories are:

- Proper Names (mainly person names)
- Toponyms; including countries, cities, regions, etc.
- Acronyms
- Events; including agreements, celebration dates, etc.
- Institutions and Political and Administrative Entities
- Jobs

On the other hand and concerning the numeric expressions, we included a category for date expressions.

Each Named Entity is tagged in an XML style. Each tag had two attributes: the ID and the type. The value of the ID attribute is numeric value unique for each Named Entity, while the type value might be one of the seven categories mentioned above.

```
<ne id="" type="">……</ne>
```

Table 1 provides examples for tagged Named Entities in the three languages.

	English	Arabic	Spanish
Proper Name	<ne type="np" id="463"> Benon Sevan </ne>,introdu-ced those documents and gave …	وقدم السيد <ne type="np" id="463"> بينون سيفان </ne>، هاتين الوثيقتين وأطلع	<ne type="np" id="463"> Benon Sevan </ne>, presentó estos documentos y ofreció
Toponym	had occurred in the northern part of <ne type="top" id="149"> Mitrovica </ne>	حدثت في الجزء الشمالي من <ne type="top" id="149">متروفيتشا </ne>	habían ocurrido en la parte septentrional de <ne type="top" id="149"> Mitrovica </ne>
Acronym	and <ne type="acro" id="163"> KFOR </ne> in good cooperation to ensure	<ne type="acro" id="163">قوة كفور </ne> لكفالة أمن جميع سكان	y la <ne type="acro" id="163"> KFOR </ne> para lograr un grado suficiente de seguridad para
Event	press statement was the <ne type="event" id="206"> United Nations Day for Women's Rights and International Peace </ne>	وبيان صحفي الموضوع المتعلق ب <ne type="event" id="206"> يوم الأمم المتحدة لحقوق المرأة و السلام الدولي </ne>	la prensa fue el del <ne type="event" id="206"> Día de las Naciones Unidas para los Derechos de la Mujer y la Paz Internacional </ne>
Institution	issues before the <ne type="inst" id="2"> Security Council </ne>	المسائل المعروضة على <ne type="inst" id="2"> مجلس الأمن </ne>	cuestiones presentadas ante el <ne type="inst" id="2"> Consejo de Seguridad </ne>
Job	nine reports by the <ne type="job" id="4"> Secretary-General </ne>	تسعة تقارير من <ne type="job" id="4"> الأمين العام </ne>	nueve informes del <ne type="job" id="4"> Secretario General </ne>
Date	<ne type="date" id="90"> 23 February 2000 </ne>	<ne type="date" id="90"> ٢٣ شباط فبراير ٢٠٠٠ </ne>	<ne type=”date” id=”90”> 23 de febrero de 2000 </ne>

Table 1. Example of tagged Named Entieies

3.3. NE Tagging

The tagging process of Named Entities was carried out in monolingual parallel modules. Most of the tasks involving NE recognition either in English or Spanish make use of the writing conventions where capitalization is applied to indicate the beginning of names of persons, places or organizations. But this rule is not applicable to Arabic, that is the reason why Arabic language supposes a challenge for NE recognition and tagging, since its writing system does not distinguish between upper and lower cases

Spanish NE Tagging: For the Spanish corpus, we used a Named Entity Tagger developed in the Computational Linguistics Laboratory of the Autónoma University.

This is a rule-based tagger, which uses a lexicon and a number of heuristics based on pattern matching. The heuristics depends mainly on the capitalization features and a small grammar for date patterns. This method achieves a high accuracy in tagging and detecting NE candidates. However, the tagger detects only two categories of Named Entities (np and date). The sub-classification of the names class is not implemented in the Tagger, so a manual validation and introduction of the different types were needed.

English NE Tagging: English Named Entities in the English corpus were tagged using an adapted version of the Spanish Tagger. The adaptation took into consideration language-dependent features and patterns. But again, a manual verification and introduction of the distinct types of Named Entities were needed.

Arabic NE Tagging: Since we lack resources for Arabic NE tagging, we had no other option than manual tagging. In spite of the laborious task of manual tagging, the fact that the corpus we are using is parallel made it possible to develop a prototype for an Arabic NE Tagger. As a starting point, this tagger takes as input the Spanish tagged Named Entities and the aligned sentence-pairs, and then it proceeds to find candidates in the corresponding Arabic sentence.

The architecture of the system is simple and it consists of three modules. The first module recognizes Arabic date patterns and compares it to the Spanish tagged dates in the corresponding aligned sentence. If they coincide, the Arabic date is tagged with the same tag of the Spanish date. The second module is a look-up module based on a bi-lingual lexicon. This module takes the tagged Spanish NE as input together with the aligned sentence pair. It looks up the bilingual lexicon and then tries to find a candidate in the Arabic corresponding sentence. The third module is a transliteration module, which tires to find a suitable candidate for Named Entities such as country names or foreign Proper Names where the Arabic language usually opts for a transliteration Given this fact, this module applies a transliteration scheme from Spanish to Arabic, trying to find valid candidates in the Arabic sentence. If a candidate is found, it is given the same tag of the Spanish NE. Although this tagger was able to recognize a big part of the Named Entities in the Arabic text, it still needs manual verification. Besides its coverage is still limited.

4. Named Entities: Patterns and Translation Across the Three Languages

4.1. General facts about the study sample

From the aligned tagged corpus, we took a sample of 300 sentences from the Spanish corpus. These 300 sentences were aligned with 307 sentences in the Arabic Corpus and 308 in the English Corpus. The difference in number of the sentences can be explained if we take into consideration common phenomena in translation such as the multi-alignments and omissions. We could consider that 1-1 alignments are the base rule to which alignments of types (1-n), (n-1), (0-1) or (1-0) are common exceptions. However, our corpus could be considered a relatively clean, free of noise corpus, but still this fact does not mean the complete absence of noise resulting from multi-alignment or omissions.

Despite the difference of NE frequencies in the sample, the whole average of Named Entities per sentence reaches 2.43. This means that each sentence might contain more

than two Named Entities. This fact reflects the nature of the text as a UN document where occurrences of Named Entities are quite frequent. Table 2 shows the frequencies of the Named Entities in the sample corpus.

	English Sample	Arabic Sample	Spanish Sample
Number of sentences	308	307	300
Total Number of NE	721	742	765
Average NE/sent	2.34	2.41	2.54
NE type "*Proper Name*"	36	39	40
NE type "*Toponym*"	158	164	167
NE type "*Acronym*"	26	11	27
NE type "*Job*"	111	123	128
NE type "*Institution*"	272	275	277
NE type "*Event*"	21	21	22
NE type "*Date*"	97	109	104

Table 2. Named Entities Distribution among the multilingual Sample

The difference in the total number of Named Entities in each corpus can be explained bearing in mind two factors:

- The different forms and patterns by which each language formulates the structure of the Named Entities
- The different strategies adopted by the translator to transfer a given Named Entity from one language to another

In fact, these two factors are strongly related and cannot be dealt with separately. The second factor is highly dependent on the first since the strategies adopted by the translator in a given case represent a subset of the applicable patterns in the Target language. In other words, given a Named Entity x, its translation y (x,y), $S=\{s_1,s_2,...s_n\}$ is the set of the strategies a translator may adopt and $P=\{ p_1,p_2,...p_n \}$ is the set of valid linguistic patterns for y given x. In this case, $S \subseteq P$.

One more fact that we would like to underline is that Named Entities are considered a s*emantic category* and not a *grammatical* one. Hence, the structure and the behaviour of Named Entities in the text are identical to those of a common noun phrase. The only difference is a semantic difference; as those Named Entities are in fact common nouns, but they have passed from common words to NE through a semantic process to refer to a certain entity, instead of a generic one. This semantic phenomenon is reflected orthographically in the use of upper case in both English and Spanish, but not in Arabic.

Regarding our analysis of patterns and linguistic structures, we will point out the most frequent cases and how their structure changes from one language to another. In this analysis, we will not follow the Named Entities classification. Instead, we will group the Named Entities in their two basic classes: names and dates. Within the names class we will focus *only* on compound Named Entities consisting in more than one lexical unit, regardless of its category if it is used to refer to an institution, a job, a toponym, etc. NE types consisting of one lexical unit are excluded in this analysis.

In the following sections we will discuss the analysis of the results in each of the three language pairs.

4.2. Data Analysis

Comparing the distribution of the Named Entities' types in the parallel corpus, the results show that the types of Named Entities are almost equally distributed in the three languages. This is expected, as it is a parallel corpus. The following graphs (Figure1, Figure 2 and Figure 3) represent the distribution percentages of types of NE in the Spanish corpus, English and Arabic corpora respectively.

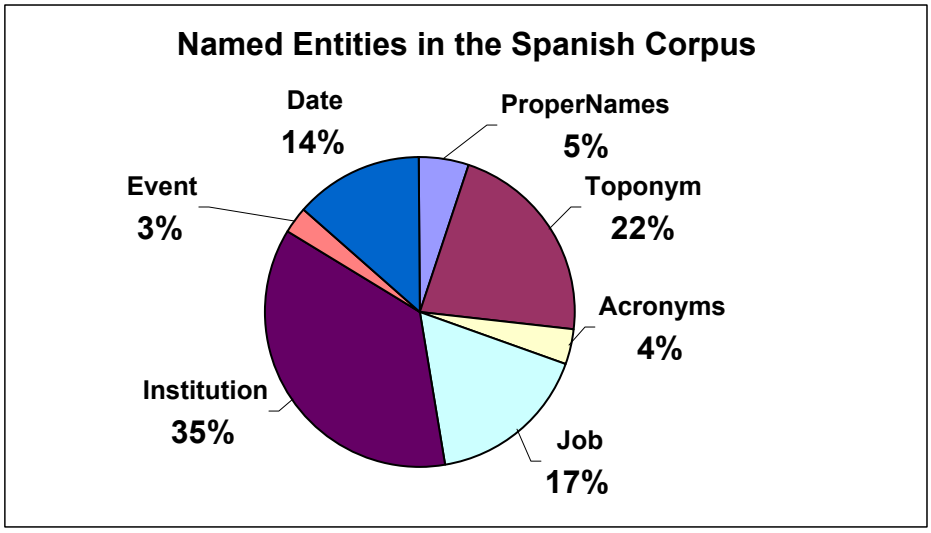

Figure 1. Named Entities Distribution in the Spanish Corpus

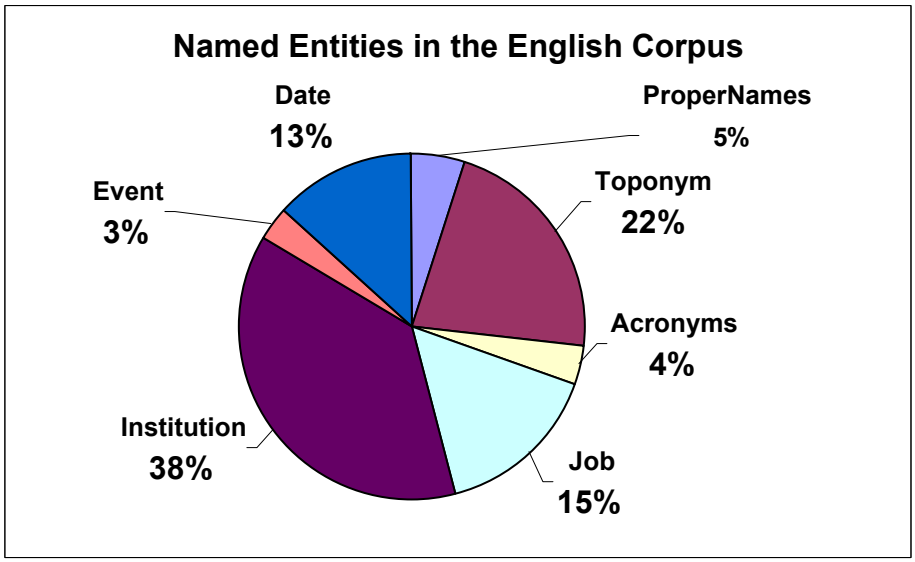

Figure 2. Named Entities in the English Corpus

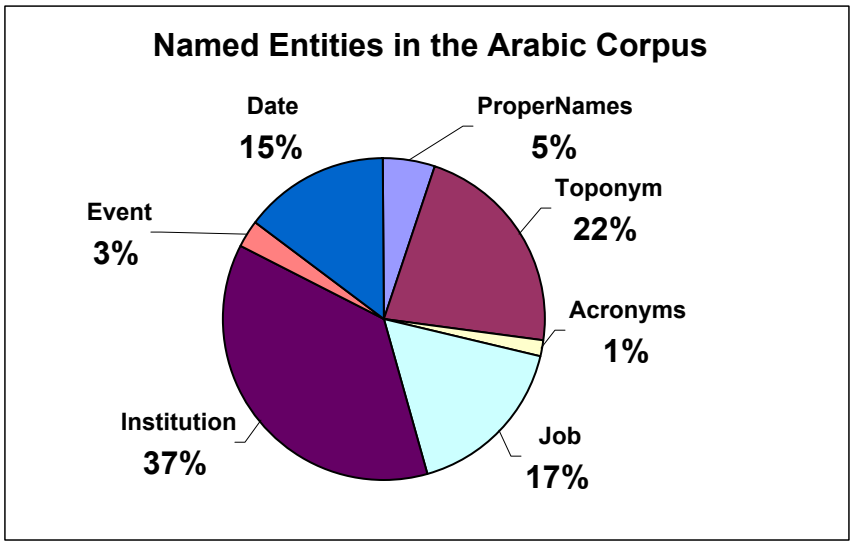

Figure 3. Named Entities in the Arabic Corpus

According to the given data, our observations could be summarized in the following points:

- Proper Names, Toponyms, Events and Dates are almost equally distributed. The main differences are observed in the types "Job" and "Institution" in the English and Spanish corpora. The "Acronyms" in the Arabic corpus only represent 1% of the total number of Named Entities

- The frequency of Named Entities of type "Job" in the Spanish corpus (17%) is higher than its corresponding type in the English corpus (15%). This is due to high frequency of the Named Entity "*Presidente*" in the Spanish corpus, especially when used in "*declaración del Presidente*". This was translated into English as "presidential statement" and so the Named Entity was substituted by an adjective.

- The higher frequency of Institution type in the English corpus is due to the fact that English tends to use shorter sentences, while Spanish merges the sentences. The merging of sentences in the Spanish language is reflected in the higher frequency of anaphors and, thus, a lower frequency of Named Entities of type Institution. For example, sentence 181 in the Spanish corpus is multi-aligned with sentences 182 and 183 in the English corpus. By comparing the number of Named Entities, we will find that the English sentences 182 and 183 have four occurrences of Named Entities, while sentence 181 in the Spanish corpus has three occurrences. The translator has opted in this case for using an anaphor within the same Spanish sentence, instead of repeating the Named Entity "*Secretaría*" of type institution/administrative entity. In English, since there are two sentences, the Named Entity "Secretariat" is repeated twice increasing the total to four.

- The same case was observed in Arabic. In our opinion, this similarity between the Arabic and the English text can be explained if we consider the probability that the Arabic text was originally translated from the English text. Although we do not have precise information about that, but it is expected that in international

organizations such as the United Nations, the main documents would be issued first in English and then translated into the rest of the official languages of the organization.

- Although the percentage doesn't reflect this observation since the difference is a very slight one, we consider it important to highlight such a feature. The exact percentage of toponyms in the Spanish corpus is 21.9%, while that of English is 21.8%. This slight difference can be attributed to the fact that the English text in very few cases opted for using adjectives instead of Toponyms as Proper Nouns, while the translator in the Spanish corpus opted for applying the Toponym as a Proper Noun. This case was observed twice. The English corpus used the adjective form "Tajik-Afgan border", while the Spanish corpus used the Proper Noun "La frontera entre Tayikistán y Afganistán", thus the Toponyms are tagged in the Spanish corpus, but not in the English corpus. The Arabic text behaved in a similar way to the English text adopting the adjective form instead of the Proper Noun.

- The Acronyms are equally distributed in the Spanish and the English corpora. However, the Arabic corpus shows a very little frequency of this type of Named Entities (only 1%). This is due to the fact that the Arabic language rarely adopts the acronyms. Instead it uses the full form of the name, or in very few cases it might opt for a phonetic transcription of the acronym. This is the case of "KFOR", which was phonetically transcripted into "كفور".

The following graph (Figure 4) shows the results of the frequencies of Named Entities in the three corpoa (English-Spanish-Arabic).

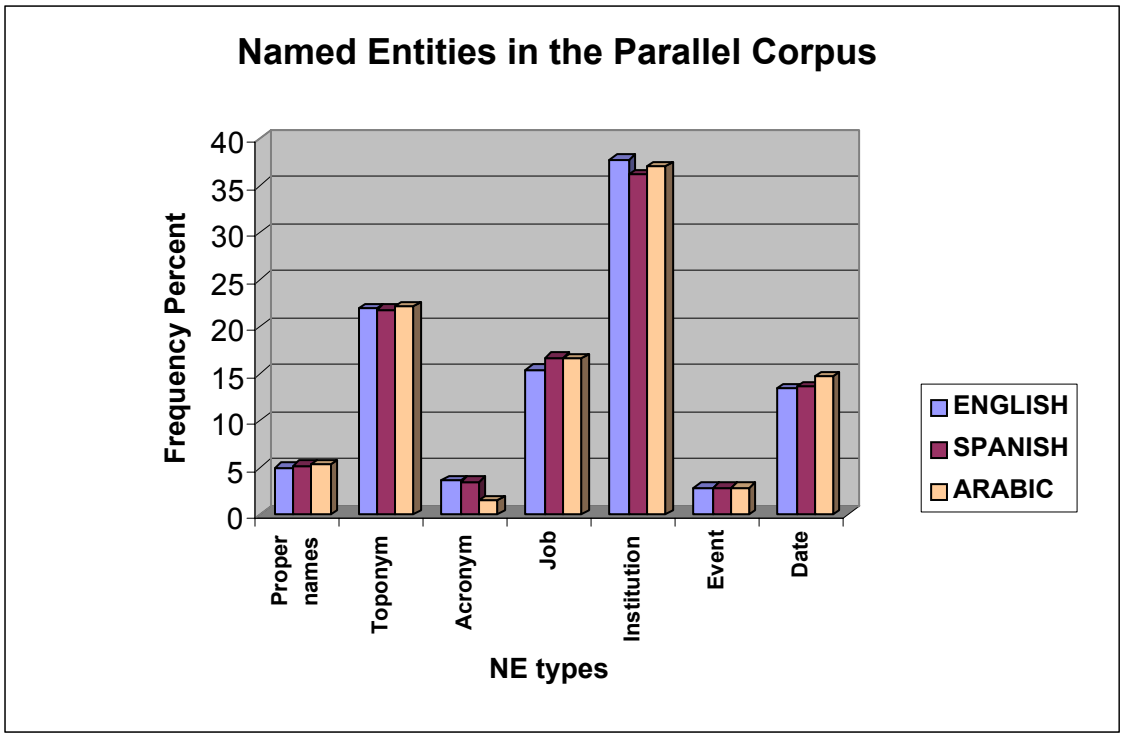

Figure 4. Named Entities in the parallel corpus

4.3. Structures and Patterns

As we mentioned before, in this section we will focus only on date patterns and the most common cases where the Named Entities adopt the form of a noun phrase.

Date patterns: The most frequent Named Entities of type "date" adopt the following patterns in English, Spanish and Arabic. We notices that the Arabic corpus applied two calendars for date expressions; the common Gregorian calendar and the Lebanese/Sirian calendar. The following table illustrates the different some from the parallel corpus.

English Date Patterns	Spanish Date Patterns	Arabic Date Patterns
NUMBER (Day) + MONTH Example: 23 February	**NUMBER + PREP (de) + MONTH** Example: 23 de febrero	**NUMBER (Day) + MONTH_LEBANESE/MONTH_GREG** Example: ٢٣شباط فبراير
NUMBER (Day) + MONTH + NUMBER (Year) Example: 23 February 2000	**NUMBER + PREP (de) + MONTH + PREP (de) + NUMBER (Year)** Example: 23 de febrero de 2000	**NUMBER (Day) + MONTH_LEBANESE/MONTH_GREG + NUMBER (YEAR)** Example: ٢٣شباط فبراير٢٠٠٠

Table 3. Date patterns English-Spanish-Arabic

Name Patterns: The following are the most common patterns of noun phrases adopted by Named Entities.

English Name Patterns	Spanish Name Patterns	Arabic Name Patterns
NOUN + ADJECTIVE Example: Secretary General	**NOUN + ADJECTIVE** Example: Secretario General	**NOUN + ADJECTIVE** Example: أمين عام
PROPER_NOUN + NOUN Example: Bangladesh Presidency	**NOUN +PREP (de) + PROPER_NOUN** Example: Presidencia de Bangladesh	**NOUN + PROPER_NOUN** Example: رئاسة بنغلاديش
NOUN + of + PROPER_NOUN Example: Government of Bangladesh	**NOUN + PREP (de) + PROPER_NOUN** Example: gobierno de Bangladesh	**NOUN + PROPER_NOUN** Example: رئاسة بنغلاديش
NOUN + of + NOUN Example: Office of the President	**NOUN + PREP (de) + NOUN** Example: Oficina del Presidente	**NOUN + (ART+NOUN)** Example: مكتب الرئيس
NOUN + NOUN Example: Security Council	**NOUN + PREP (de) + NOUN** Example: Consejo de Seguridad	**NOUN + (ART+NOUN)** Example: مجلس الأمن
ADJECTIVE + NOUN Example: Special Envoy	**NOUN + ADJECTIVE** Example: Enviado Especial	**NOUN + ADJECTIVE** Example: مبعوث خاص

Table 4. Name patterns English-Spanish-Arabic

These common patterns can combine together to form more complex units such as "Special Envoy of the Secretary General". In such cases the resulting pattern is the sum of its constituents "**(ADJECTIVE+NOUN)+ of + (NOUN+ADJECTIVE)**". In Spanish, the pattern "**NOUN+ADJECTIVE**" is applied for both constituents and the final result is "**NOUN+ADJECTIVE**" PREP(de) "**NOUN+ADJECTIVE**" "*Enviado Especial del Secretario General*". The Arabic will adopt the same strategy where the result is the sum of combination of the two patterns "**NOUN+ADJECTIVE**" PREP(ل) "**NOUN+ADJECTIVE**" and thus the translation is"**المبعوث الخاص للأمين العام**".

Analyzing the most common structures of noun phrases adopted by the Named Entities, we could observe that in some cases the human translator has more than one possibility. For example, if we took the Spanish text as a starting point and the English as a Target language, the Spanish pattern **"NOUN + PREP (de) + PROPER_NOUN"** translated into English might adopt one of two possible patterns; "**PROPER_NOUN + NOUN"** or **"NOUN + of + PROPER_NOUN"**. The two solutions are valid and the decision of the human translator in this case reflects the translation strategy adopted.

5. Conclusion and Future Work

Studying Named Entities in a multilingual context is a novel approach. However, we consider the present study a starting point that opens new dimensions and reveals more facts about the underlying linguistic features of such semantic categories. The results of this study might give some useful insights for a wide range of fields.

Fields like Information Extraction, Natural Language Processing and Computational Linguistics might benefit from these results in order to enhance their systems and models. At the same time, these results could give some clues to translation studies for a better and a deeper understanding of the underlying translation strategies across three completely different languages.

On the other hand, in the field of Language Teaching and Language Acquisition tagged parallel corpus proved to be a valuable resource, as it provides the teacher and students with real-life examples and introduces them to the actual use of the language.

Another application which would benefit directly from this study is the development of Example-Based Translation Systems, since the main idea behind such systems relies on analysing the different patterns and how they are translated from one language into another.

For future work, we would consider extending the size of the aligned tagged corpus, this would provide us with more linguistic facts and more resources, which would help us take a further step towards the comparable corpora. Also the Named Entity Tagging would be a starting point for a wider semantic tagging on a multilingual level. On the level of application, a study of how such a parallel corpus could be adapted for pedagogical purposes is one of our short-term goals.

References

Books:

Baker, M. (ed.) (1998) *Encyclopedia of Translation Studies* (London-New York: Routledge).

Laviosa, S.(2002) *Corpus-based Translation Studies: Theory, Findings, Applications*. (Amsterdam/New York, NY).

Melamed, I. D. (2001) Empirical Methods for Exploiting Parallel Text (Cambridge/London: MIT Press).

Articles in Book:

Baker, M. (1993) Corpus Linguistics and Translation Studies: Implications and Applications, in Baker, M., Francis, G., & Tognini-Bonelli, E. (eds.). *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 233-250.

Articles in Journal:

Al- Onaizan, Y. and Knight, K. (2002) Machine translation of names in Arabic text. Proceedings of the ACL conference workshop on computational approaches to Semitic languages.

Al- Onaizan, Y. and Knight, K. (2002) Translating Named Entities Using Monolingual and Bilingual Resources. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* Philadelphia, July 2002, pp. 400-408.

Baker, M. (2004) A corpus-based view of similarity and difference in translation, in Journal of Corpus Linguistics vol. 9, no. 2, pp. 167-193 (John Benjamins Publishing).

Grishman, R. and Sundheim, B. (1996) Message Understanding Conference - 6: A Brief History. *COLING-96*.

Hasegawa, T. Sekine, S. Grishman, R. (2004) Discovering Relations among Named Entities from Large Corpora. In *Proceedings of the Annual Meeting of Association of Computational Linguistics (ACL 04)*; Barcelona, Spain.

Larkey, Leah, Nasreen AbdulJaleel, and Margaret (2003) What's in a Name?: Proper Names in Arabic Con-nell. Cross Language Information Retrieval, CIIR Technical Report, IR- 278.

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

Morarescu, P. and Harabagiu, S.(2004) NameNet: a Self-Improving Resources for Name Classification. *LREC-2004* , pp.717-720

Samy, D., Moreno Sandoval, A. and Guirao, J.M (2004) An Alignment Experiment of a Spanish-Arabic Prallel Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR 2004),* Cairo, Egypt, pp.85-89.

Stalls, B. and Knight, K. (1998) Translating Names and Technical Terms in Arabic Text. *COLING/ACL Workshop on Computational Approaches to Semitic Language*s. Montreal, Quebéc.

On-line publications:

Somers, H. (2001) Bilingual Parallel Corpora and Language Engineering (*Anglo Indian Workshop "Language Engineering for South Asian Languages*" *LESAL*, Mumbai. Available on-line from: http://www.emille.lans.ac.uk/lesal/somers.pdf (accessed April 10th, 2005)

Sekine, S. (2004) Named Entity: History and Future. Available on-line from http://cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf (accessed April 10th, 2005)

Laviosa, S.(2000)A Resource for Studying What Is "in" and "of" Translational English. Available on-line from http://www.art.man.ac.uk/SML/ctis/events/Conference2000/corpus1.htm (accessed June 10 th, 2005)

Laviosa, S.(2000) The Corpus-Based Approach: A New Paradigm In Translation Studies. Available on-line from http://www.erudit.org/revue/meta/1998/v43/n4/003424ar.html (accessed May 17 th, 2005)