Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic-Spanish-English)

Doaa Samy[1,2] and Ana González-Ledesma[2]
[1] Cairo University, Egypt
[2] Laboratorio de Lingüística Informática-Universidad Autónoma Madrid, Spain (LLI-UAM)
[1] Cairo University, Main Campus, Spanish Dept. Faculty of Arts, 12613, Giza, Egypt.
[2] Laboratorio de Lingüística Informática, Dpto. de Lingüística, Facultad de Letras, Cantoblanco 28049 Madrid
E-mail: [1] dsamy@cu.edu.eg, [1, 2] {doaa,ana}@maria.lllf.uam.es

Abstract

Discourse structure and coherence relations are one of the main inferential challenges addressed by computational pragmatics. The present study focuses on discourse markers as key elements in guiding the inferences of the statements in natural language. Through a rule-based approach for the automatic identification, classification and annotation of the *discourse markers* in a multilingual parallel corpus (Arabic-Spanish-English), this research provides a valuable resource for the community. Two main aspects define the novelty of the present study. First, it offers a multilingual computational processing of discourse markers, grounded on a theoritical framework and implemented in a XML tagging scheme. The XML scheme represents a set of pragmatic and grammatical attributes, considered as basic features for the different kinds of discourse markers. Besides, the scheme provides a typology of discourse markers based on their discursive functions including hypothesis, co-argumentation, cause, consequence, concession, generalization, topicalization, reformulation, enumeration, synthesis, etc. Second, Arabic language is addressed from a computational pragmatic perspective where the identification, classification and annotation processes are carried out using the information provided from the tagging of Spanish discourse markers and the alignments.

1. Introduction

Pragmatics is usually defined as the study of how language is used. In the language use, context plays a key role in the interpretation of statements. That is the reason why Pragmatics is concerned, among other topics, with *Inference*. Through this mental process, humans can obtain information that is not actually present in the utterance or statement at hand (Sperber & Wilson, 1986).

From a computational point of view and according to Jurafsky (2002): "four core inferential problems in pragmatics have received the most attention in the computational community: REFERENCE RESOLUTION, the interpretation and generation of SPEECH ACTS, the interpretation and generation of DISCOURSE STRUCTURE AND COHERENCE RELATIONS, and ABDUCTION". In the present study, we focus on the discourse markers as key elements in guiding the inferences of the statements. We present a resource for the community that addresses aspects concerning DISCOURSE STRUCTURE AND COHERENCE through the automatic identification, classification and annotation of the *discourse markers* in a multilingual parallel corpus (Arabic-Spanish-English).

1.1. Motivation

Discourse markers have been subject of different studies in the field of computational pragmatics and natural language processing. especially in applications concerned with the detection of document structure for automatic summarization or for the interpretation and generation of speech acts in speech corpora and dialogue systems (Kawahara & Hasegawa, 2002). However, most of these studies focused on the automatic ambiguity resolution of only certain discourse markers (Zufferey & Popescu-Belis, 2004) or in the classification of the markers in monolingual corpora. Moreover, the majority of these studies deal with the ambiguity from a merely technical perspective which doesn't study in depth the *linguistic* ambiguity in discourse markers. This ambiguity can be categorial, discursive or both.

On the other hand, the state-of-the-art reveals a lack of studies considering discourse markers in Arabic from a computational perspective. To our knowledge, they have been briefly mentioned in the annotation tools provided by the LDC for the annotation of Arabic speech corpora (Strassel & Walker, 2004), while the rest of the studies adopted a completely theortical linguistic point of view (Sarig, 1995).

Given all the above mentioned facts, the novelty of our study lies in two main aspects. First, this work offers a comprehensive multillingual treatment of the discourse markers. Second, it addresses the written Arabic language from a computational pragmatic perspective.

1.2. Work Outlines

The present study is organized as follows. After this introductory section, in the second part, we explain the guidelines defining our theoretical pragmatic framework. Based on this framework, in the third section, we describe the typology adopted in the classification of the discourse markers and how it is reflected through the PRAGMATEXT, the XML annotation model for pragmatic information. The fourth section discusses the characteristics of the corpus and the methodology used for the automatic detection, classification and annotation of the discourse markers in the three languages. The approach takes into account two types of challenges. First, the levels of ambiguity in discourse markers, especially in the Spanish considered as the starting point for tagging the Arabic corpus. Second, the different technical strategies

adopted to identify and tag the discoure markers in the three languages, Spanish, English and Arabic. Results of Arabic tagging are included.
Finally in the last section, conclusions are drawn and future work is outlined highlighting some of the possible applications that can benefit from such a resource.

2. Theoretical Framework

Studies in Pragmatics revealed that the communication process is carried out at the inferential level, i.e., the communication process is no longer regarded as a process of encoding and decoding the information, but as a process where the interpretation of the world of the speaker is encoded, transmitted, decoded and finally interpreted again by the interlocutor. According to this model, each linguistic utterance has an argumentative charge responsible of raising certain inferences in the interlocutor. Discourse markers play an important role in guiding these inferences and, therefore, tasks concerning their identification and annotation become indispensable for language processing in order to reach to a complete understanding of the underlying meaning of each utterance.
The theoretical framework adopted proposes a correspondence between the cognitive, socio-cultural level on one had and the linguistic levels, on the other hand. Following this premise, a type of reasoning is reflected through a discursive operation. This operation is linguistically encoded through discourse markers. However, the linguistic encoding of discourse markers varies from one language to another, since each language adopts different morphological, syntactical or lexical strategies to express the discursive operation.
In other words and according to the theoretical framework implemented in the PRAGMATEXT annotation model, a discourse marker implies an inference, considered as a cognitive universal phenomenon. However, such phenomenon is modelled in a different way by each group of users; expressed by different grammatical strategies in each language and, thus, materially realized taking different linguistic forms.
Given these cognitive and linguistic facts, a computational approach dealing with discourse markers should take into consideration the following theoretical challenges:
1) The lack of consensus regarding the classification of what is considered as a discourse marker and what is not and, consequently, what is the definite set of discourse markers in a given language.
2) The categorical (grammatical) ambiguity; a discourse marker could be at the same time an adjective, adverb or a whole phrase. For example, in Spanish, "bueno" (*well*) could be a discourse marker in some cases and an adjective in other cases. In the same way, the Arabic "حسناً" (*well*) could also be a discourse marker or an adjective. Similar is the case of the English "well" which can be a discourse marker or an adverb.
3) The syntactic ambiguity, i.e., some markers operate at the sentence level and at the phrase level. For example, the conjunction "and" in English and its equivalents in Spanish "y" and "و" in Arabic.
4) The discursive ambiguity; one discourse marker can have different discursive functions. For example, the Arabic "كما" and its English equivalent "*as*", both can play the role of a marker of concretion. However in Arabic, it could also be a co-argumentation marker and in English, it could be a marker of cause as well.
5) Idiomatic expressions, i.e., a decision has to be taken in case two discourse markers appear simultaneously; either to be considered as two different discourse markers or as a multi-word discursive unit. For example, the English "as" appears in "as well as" and "as to". In the first case it is a co-argumentative marker, while in the second it is a topicalization marker.
For the first challenge, more time and more work is need to be directed to the study of discourse markers in different types of corpora and in the different language, both in the spoken and the written register. Accordingly, the present work is an initiative and a step forward to reach this goal.
Resolving the grammatical and syntactic ambiguity, representing the second and the third challenges, requires disambiguation methods, either by means of contextual rules or statistical methods. However, for the latter approach, the availability of resources with both POS and tagged discourse markers is the main obstacle. For the present study, contextual disambiguation rules, using information from the POS tagging, were applied only to the Spanish. Once the Spanish discourse marker is identified, we search both the English and the Arabic corpora for an equivalent marker.
Regarding the discursive ambiguity and the discourse markers formed up from more than one marker, the attitude of theoretical linguists varies from that of the computational linguists. Theoretical linguists opt for increasing the discursive values for each marker. Computational linguists and computer scientists, on the other hand, prefer a well-defined and stable set of discourse markers applicable to wider range of domains and registers.
Considering the above mentioned challenges, the proposed annotation model pretends to reach a compromise between the theory and the computational practice by decreasing the ambiguity, but at the same without losing the details provided by the Pragmatic theories. The following section describes the annotation model.

3. PRAGMATEXT: A Model for Pragmatic Annotation

PRAGMATEXT is a model for pragmatic annotation, designed mainly for tagging phenomena related to the truth of the statements in spoken corpora (González-Ledesma, 2007).
According to PRAGMATEXT, pragmatic information is marked up by a `<PI>` tag. This tag can be used to annotate different pragmatic phenomena on the sentence level in a given corpus. However, in this case, it is used to tag the discourse markers.
Predicate Logic and Formal Semantics revealed that statements of any language could be true or false. Such an interpretation is modified by the subjective perception of the reality. Such perception can be defined through a number of phenomena responsible of modifying the interpretation. These phenomena are represented in PRAGMATEXT model through a set of attributes assigned to each tag of Pragmatic Information `<PI>`. The following are the seven phenomena affecting the truth of the statement:
1) Emotional Language `[ED]`, including judgements with positive or negative evaluation, expressions of surprise or exclamations and interjections.

2) Discursive Relations [DR], referring reasoning strategies, the argumentative charge of a statement or the discursive operations concerning the verbalization of certain mental operations such as: generalization, concretion, co-argumentation, contra-argumentation hypothesis, condition, cause, concession, reformulation, topicalization, how, synthesis, time, purpose, etc.
These kinds of discursive relations concerned with the verbalization of mental operations define the typology adopted for our corpus annotation.

Discursive Function	**Example**		
	ES	**AR**	**EN**
Topicalization	Respecto de	فيما يتعلق	Regarding
Generalization	en general	بصفةعامة	In general
Co-argumenation	también /y	أيضا/و	as well as/and
Co-argumenation1	En primer lugar	في المقام الأول/ بداية	first
Co-argumenation2	Por otra parte	ومن ناحية أخرى	On the other hand
Co-argumenation3	finalmente	أخيرا	finally
Contra-argumentation	pero	لكن	but
Concretion	en particular	خاصة لاسيما	in particular
Concession	aunque	بالرغم من	although
Cause	porque	لأن/إذ أن	because
Condition	una vez que	بمجرد	once
Hypothesis	si	إذا كان	if
How	de esta forma		in this way
Purpose	Para que	لكي	in order to
Time	tan pronto como	بمجرد	as soon as
Reformulation	a saber	و هي	namely
Option	o	أو	or
Simultaneity	Al tiempo que	في حين	while
No	Posible-mente	قد	possibly

Table 1. Typology of Discourse Markers

3) Discursive Modality [MOD] reveals the speaker's commitment regarding the truth of the statement. There are three types of modality: attenuation, intensification and interaction. For example, in Spanish, "¿Me entiendes?" (*Is it clear/Do you understand?*).
4) Evidentiality [EVI] refers to the source of information on which the speaker grounds his judgement regarding the truth of the statement. For example, sources of information could include written or oral sources, senses such as vision, other persons, etc. For example, "aparentemente" (*Apparently*), "según X" (*According to X*), etc.
5) Metaphor [MET]: this phenomenon has to do with the semantic fields of both the source and the target domains. For example, the body is the source domain in the marker "on the other hand" while the discourse is the target domain.
6) Speech Acts [SA]
7) Deixis [DEX]: defines deictic references to the context. Deixis are either social, such as, "hombre", (*man, bud, dude*) or textual, such as "anteriormente" (*previously/before*).
In addition to the above mentioned attributes, the <PI> tag includes other relevant attributes such as:

a) An identification number [ID]
b) The discursive position in the sentence (initial, intermediate or final)
c) The grammatical category [GC] (POS or type of phrase)
d) The lemmas of the constituent elements [Lema]
e) The level of idiomacity [FU] (if it is a collocation or a locution)
f) The pragmatic category [Range] (either operator or connector)

The following is an example of the tag assigned to the discourse marker "por ejemplo" (*For example*):

```
<PI ID="1"
    Lema1="por"    Lema2="ejemplo"
    GC="Prepositional_Phrase"
    DP="2"
    Range="operator"
    FU="Loc"
    MET="No"
    DR="Concretion"
    ED="No"
    MOD="No"
    EVI="No"
    SA="No"
    DEX="No"
>
por ejemplo
</PI>
```

For the annotation of the discourse markers in the present multilingual corpus, PRAGMATEXT was adopted, but with minor modifications, taking into consideration that it is a written corpus and some of the phenomena contemplated are not applicable or they will usually be assigned a negative value. For example, the "modality" attribute in our corpus is assigned a negative value, as it is a written formal discourse where the writer (or in this case the translator) does not reveal any commitment towards the truth of the statement.
On the other hand, PRAGMATEXT initiative, though language independent as it deals with universal cognitive phenomena, was mainly applied on the Spanish language. That is the reason why we use the Spanish corpus as a staring point for the annotation in English and Arabic in this study. At the same time, conclusions drawn from such approach help to evaluate the feasibility and the adequacy of the annotation model to deal with different languages.
Moreover, the set of features provided through the attribute-value pairs, helps in resolving part of the theoretical discursive ambiguity. Besides and from a computational perspective, it is useful in training statistical and machine learning models.
Despite the fact that at this stage of the research no statistical models nor learning techniques are applied, but it is a first step to build a reliable resource that could be used in applying such approaches in the future.

4. Discourse Markers Annotation

Once defined the framework, we proceed with the application where we describe the data used in our experiment and the main challenges encountered.

4.1. Resources and System Architecture

4.1.1. Resources

Our experiment was carried out using the following resources:

- A part of a trilingual parallel corpus (English-Spanish-Arabic), formed up from UN documents, aligned on the sentence level and tagged on the POS level (Samy et al., 2006). The part of the corpus used for the experiment is made up from 40,000 words in Spanish and their equivalent in English and Arabic. Table 2 shows basic information concerning the corpus in use.

	Spanish	Arabic	English
No.of Tokens	39.496	26.179	32.893
No. of Sentences	1179	1173	1182

Table 2. Corpus Information

- A lexicon of Spanish discourse markers with their different attributes as indicated in section 3.
- A bilingual lexicon of Spanish-English discourse where Spanish is used as a source language and the equivalent English discourse markers are provided to the source markers.
- An automatically translated bilingual lexicon of Spanish-Arabic discourse markers [1]. The machine translation provided the Arabic discourse markers equivalent to the Spanish markers where as in the previous case, Spanish is used as a source language.

It is important to highlight that the in both bilingual lexicons, the translation is not on a one-to-one basis. One Spanish discourse markers may have more than an equivalent in the target language and two or more Spanish discourse markers might have one equivalent in the target language.

4.1.2. System Architecture

A rule-based approach is adopted for the automatic identification and classification of the discourse markers. The implemented algorithm is based on the following hypothesis:

In an aligned pair of sentences, if a discourse marker appear in a sentence in the Spanish corpus, it is most probable that the corresponding English and Arabic sentences contain a discourse marker.

Our approach consists of two phases.

Phase 1 is formed up from a monolingual module for the annotation and disambiguation of Spanish discourse markers.

Resources used in this module are: the monolingual Spanish corpus and the monolingual Spanish lexicon of discourse markers. The output is the Spanish corpus with discourse marker annotated and disambiguated.

[1] Google translate: Spanish-English, then English-Arabic BetaWikiled Online Dictionary (Spanish-Arabic): http://www.wikiled.com/spanish-arabic-Default.aspx

Phase 2 consists of two modules: the annotation of English discourse markers and the annotation of the Arabic discourse markers.

Each module in phase two has as an input the bilingual lexicon and the set of sentence alignments for the language pair in concern, i.e., Spanish-English and Spanish-Arabic. The output of each module consists of the corpus of the target language tagged with the corresponding discourse markers inheriting the features and attributes of the source discourse marker (in Spanish).

4.2. Annotation of Discourse Markers in Spanish

Information provided from the input monolingual Spanish lexicon of discourse markers is searched in the Spanish corpus. Occurrences of each discourse marker in the lexicon are tagged with the necessary information in the <PI> tag. However, discourse markers represent certain level of ambiguity as mentioned in section 2. To resolve these ambiguities, the discourse markers in the lexicon are classified into four types:

1) Non-ambiguous discourse markers
2) Categorial ambiguous
3) Discursive ambiguous
4) Categorial discursive ambiguous

According to the type of discourse markers, different strategies are followed when assigning the pragmatic information through the <PI> tag.

1. If the discourse marker is non-ambiguous, it is automatically tagged.

2. For categorical ambiguous markers, context rules are implemented considering two types of features:

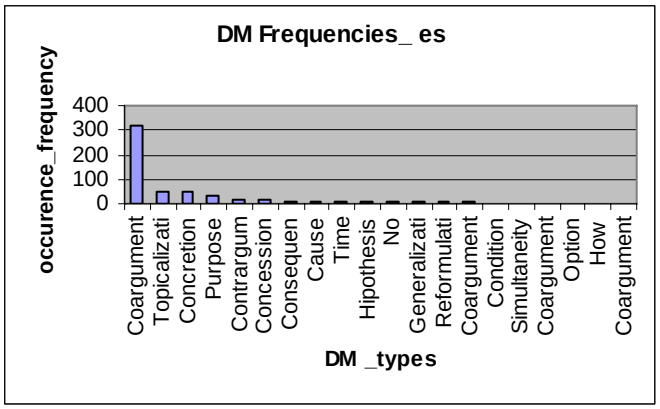
- prosodic features reflected through the punctuation; and
- the position of occurrence within the sentence (inter-sentential segment).

For example, the discourse marker "finalmente" (finally) could be ambiguous in some cases if used within a sentence as an adverbial complement describing a verb phrase and in this case would be synonymous of "por fin" {e.g. "han llegado finalmente a un acuerdo" *they finally reached an agreement*)}. However, it is considered as a discourse marker if it occurs at the beginning of a sentence followed by a comma {e.g. "Finalmente, vamos a tartar el aspecto 2" (*Finally, we are doing to discuss aspect 2*)}.

Ambiguous discourse markers of type 3 and 4 represent a small percentage of the cases detected in the corpus and they were disambiguated manually. This is due to the fact that it is a written formal corpus. However, in case of larger corpora or spoken corpora, statistical and machine learning techniques can be used to handle such ambiguities.

The Spanish annotation module was able to tag 558 occurrences of discourse markers with their different discursive types in 418 sentences of the whole corpus (1179 sentences). In this way, discourse markers appear in almost 34% of the total number of sentences. The 558 occurrences represent 83 unique discourse markers where the co-argumentation markers achieve the highest frequency with 321 occurrences {e.g. "y", "también", "así como" (*and, also, as well as*)}, followed by topicalization with a total of 47 occurrences {e.g. "respecto a", "en cuanto a" [*regarding, concerning, as to*]} and concretion with a total of 47 occurrences {e.g. "en particular", "a saber" (*particularly, namely*)}. Results of the frequencies of different types of Spanish discourse markers are shown in Graph 1.

The final output of this module consists of the Spanish corpus with the discourse markers tagged, together with their types, their grammatical categories and their attributes. Information regarding the grammatical category of each discourse markers is automatically assigned after keeping track of the sequence POS of the constituents lemmas. For example, the discourse marker "en particular" (*particularly*) is assigned the value of Prepositional Phrase for the Grammatical Category Attribute. This value is detected by keeping track of the sequence of the POS of the constituents' lemmas, which in this case is Preposition+ Adjective.



Graph 1. Frequencies of DM in the Spanish Corpus

4.3. English and Arabic Annotation Module

The second phase consists of tagging the English and Arabic discourse markers.

The two modules for English and Arabic discourse markers' annotation make use of two sets of the parallel corpus, the Spanish-English and the Spanish-Arabic, respectively. In addition, two more inputs are used: the corresponding bilingual lexicon and the set of sentence alignments for each of the indicated sets in the parallel corpus. The followed algorithm implements two search types:

- Lexicon search
- Heuristic search

For the lexicon search, the procedure is as follows:

- For each sentence in the Spanish corpus, the module extracts the tagged discourse markers.
- Each discourse marker is looked up in the bilingual lexicon and its equivalent discourse markers are retrieved.
- Through the set of alignments, the aligned target sentence is searched for any occurrence of the equivalent discourse markers retrieved from the lexicon.
- If an equivalent is found it is annotated by the <PI> and it is give the same ID.
- Pragmatic attributes are inherited from the source Spanish tag.
- Attributes regarding lemmas are assigned making use of the information provided by the POS tags. Grammatical Category is detected by the sequence of the POS tags of the constituents' lemmas in each of the target languages.

Heuristic search is applied in case the lexicon search fails to find a corresponding candidate in the target language or in case no translations are provided for the Spanish source discourse marker.

Based on basic observations from the corpus regarding the co-textual features of the occurrence of discourse markers, the heuristic search considers two main factors: the inter-sentential segment position and the delimiters used. Delimiters are represented through the punctuation marks{/./, /:/, /، /}, which in turn constitute the conventions adopted by the writing system to reflect the prosodic features of the language

Examining the data in the corpus, we noticed that the different corpora adopt, more or less, similar prosodic features to indicate the discourse markers. Thus, this information is used in searching a candidate segment for discourse markers within the target sentence.

In Figure 1, we notice that following the heuristics indicated by the position of the inter-sentential segments determined by the punctuation marks as delimiters, can help locate the candidate as it is the third segment in the sentence in the three language. Although, in Spanish the third inter-sentential sentence is marked by different delimiters: [/,/ discourse marker/:/], we do not take into consideration the exact type of delimiter, we only detect the presence of the delimiter.

Arabic	English	Spanish
ووصولا إلى هذه الغاية، عدد أربع نقاط أساسية، **وهي بالتحديد**، دعم المجلس القوي للسعي إلى إيجاد حل إطاري، والمشاركة النشطة لـ دول الإقليم في تحقيق هذا الهدف، والإشارة الواضحة إلى أن الحل الحقيقي ينبغي أن يفي بالحد الأدنى من متطلبات جميع الأطراف، لا أن يحقق المطالب القصوى لأي منهم، والتوصل إلى اتفاق يكون عنصرا أساسيا في إطار تسوية أوسع تشمل المنطقة ككل	To that end, he listed four benchmarks, **namely**, the firm support of the Council for the search for a regional solution, the active participation of the States of the region in achieving that goal, the clear indication that a true solution should meet the minimum requirements of all parties, but the maximum demands of none, and an agreement that was firmly embedded in the context of a broader settlement encompassing the region as a whole.	A tal fin , enunció cuatro condiciones , **a saber** : el firme apoyo del Consejo a los esfuerzos por lograr una solución regional , la participación activa de los Estados de la región en la realización de ese objetivo , una clara indicación de que toda solución verdadera debía satisfacer los requisitos mínimos de todas las partes , pero no las exigencias máximas de cualquiera de ellas , y un acuerdo firmemente arraigado en el contexto de una solución más amplia que abarcase a la región en su totalidad .

Figure 1. Example of similarities in occurrence position of discourse markers in the multilingual corpus

The next steps are the same as the Lexicon search. The candidate segment is tagged in the target corpus inheriting the types of its corresponding source marker. Lemmas and grammatical category is assigned making use of the POS tags provided already in the corpus.

In this respect, we have to point out that Arabic implies certain challenges either due to the morphological features or the orthographical features. An example of morphological challenges, is the presence of clitics, where one token is formed up from more than POS unit. In this case, peeking track of the sequence of constituents does not only apply on a word/token-basis, but it has to take into consideration constituents of each token. Nevertheless, the availability of this information through the POS annotation level makes our task easier.

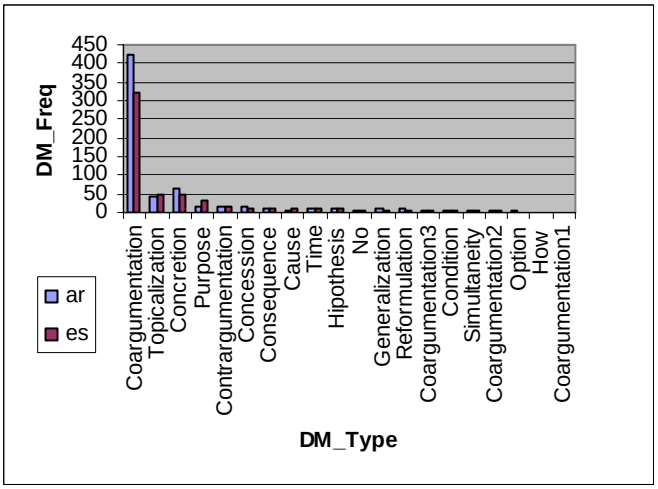
As to orthographic features, Modern Standard Arabic writing conventions usually do not separate between the conjunction "**و**" (*and)* and the following token. This fact causes much noise, as it is difficult to distinguish between the character "**و**" when it appears at the beginning of the word or when it appears as a conjunction. Again, this type of difficulties could be solved from previous annotation levels as the POS.

An example representing both challenges is the Arabic equivalent of the discourse marker "namely" ("وهي"). The Arabic token is formed up from a noun phrase formed up from two grammatical units:

و هى	-->	CONJUNCTION+PRONOUN

The Arabic annotation module was able to detect 664 occurrences of discourse markers in 427 sentences of the whole corpus (1173). The 664 occurrences represent 223 unique discourse markers. 449 of the 664 detected are correct with a precision of 67.6% on a monolingual basis. A bilingual evaluation would consider the number of correctly identified translation equivalents given the source Spanish markers. Following this criterion, the applied approach was able to detect 449 correct equivalents for the 558 Spanish discourse markers achieving a precision of 80.4%. However, the evaluation of these results is still at an early stage. For a complete evaluation, a golden standard is needed where the whole corpus should be manually validated for discourse markers. At this preliminary stage of research, we are still testing our approaches. That is why, we only evaluated the precision.

The results of the automatic detection of the Arabic discourse markers are shown in Graph2.



Graph 2. Frequencies of DM in Arabic and Spanish

As observed in the graph, discourse markers with a co-argumentative role are the most frequent in both Spanish and Arabic corpora. However, the high frequency of the discourse marker "y" (and) occurring 201 times in Spanish from a total of 321 co-argumentative markers caused much noise when trying to find its Arabic equivalent "و". In Arabic, "و" is ambiguous due to the writing conventions, as we mentioned previously. Besides, Arabic language, for stylistic reasons, use this discourse marker with a high frequency compared to other languages. In this way, results reveal that almost the majority of the sentences start with "و" (215 occurrences) increasing the total number of co-argumentative markers to 422 compared to the Spanish.

Errors in locating the corresponding discourse markers in the target language are due to different reasons. First, in the translation process, the one-to-one relation is not always applicable. In some cases, the translator adopts a different strategy of the language to express the same content without using a discourse marker. Such cases affect the automatic detection of the discourse markers and, thus, the overall precision of the system.

The main reason behind the high recall is the heuristic search. In many cases, the position of the segment where the discourse marker occurred, changes from Spanish to Arabic. Besides, in many cases, there were omissions of explicit discourse markers and the translator opted for nominalization or the use of verb phrases to express discursive functions such as purpose. For example, the Spanish phrase "hizo un llamamiento para que se respetara" is translated in Arabic as "وطالب باحترام" equivalent to (called for their safety to be respected).

It is also important to point out that the use of punctuation marks in Arabic writing conventions is not a well-defined practice, i.e., its application and use are not well-normalized. As shown earlier in Figure 1, this type of texts represents some similarities in using the punctuation marks in the different languages. These similarities, however, are the result of adopting western conventions from the source text. Moreover, results obtained show that it is difficult to generalize the observation stated in Figure l, since there is still a lack of normalization in using the punctuation marks in Arabic. This is one of the main reasons responsible of increasing the error rate in the heuristic search module and, thus, affecting the overall precision for the Arabic annotation.

English language, on the other side, achieved higher precision, compared to the Arabic, due to the higher accuracy in the heuristic search. Similarities between Spanish and English regarding the position of occurrence of discourse markers and the use of punctuation marks increased rates of the heuristics search affecting the overall precision. For the 558 Spanish source markers, the applied approach correctly detected ### corresponding markers achieving a precision of ##%.

For the final output, discourse markers are highlighted using different colours indicating their types. Such visual effects make it easier for the user to navigate through the text following the structure of the document. A snapshot of the aligned Spanish-Arabic corpus with the discourse markers highlighted is included in Appendix A.

5. Conclusions and Future Work

This study is a first attempt to bridge the gap between the knowledge provided by the theoretical Pragmatics and its implementation in the Computational Pragmatics in a multilingual context. Starting from a theoretical framework, the annotation model tries to encode the different pragmatic phenomena through a set of features and attributes. Since this model addresses universal cognitive features, it is language independent. However, it is the first time to apply it on other languages than the Spanish. In this way, we were able to prove the feasibility and the adequacy of the model to other languages.

Though each language has its own ambiguities and challenges, using a generic model and a parallel corpus, show that once these ambiguities are resolved in one language, it is feasible to apply them to the corresponding texts in other languages. Using parallel corpus in disambiguation is a common practice especially in word sense disambiguation. However, here we apply it in the domain of discourse markers.

On the other hand, the experiment carried out in this study reveals many useful facts concerning the specific strategies each language adopts to encode the discursive structure. These facts help the community build a better analysis and understanding tools for each language. Besides, results are also valuable to discover some facts related to the translation process and the strategies adopted to encode the discursive functions. In this way, we believe that providing such a resource for the community could help both the language specialists in defining and analyzing the pragmatic phenomena, and the computer scientists in developing systems which could offer a better understanding and analysis of natural language texts and its structure. These systems might have their applications in a wide range of fields such as: automatic summarization, machine translation, information extraction, etc.

Finally, as future work, we plan to develop a statistical disambiguation module to enhance the detection and disambiguation of discourse markers and to test the validity of this approach on more languages.

Moreover, a direct outcome of this research consists in developing different subsets of monolingual discourse markers, especially for Arabic, with their associated attributes. These sets would be used together with statistical disambiguation to detect and classify discourse markers in monolingual texts.

6. Acknowledgements

This research has been partially funded by the Spanish Ministry of Education under the grant TIN2007-67407-C03-02

7. References

González Ledesma, A. (2007): PRAGMATEXT: Annotating the corpus C-ORAL-ROM with Pragmatic Knowledge. In *Proc of Corpus Linguistics 2007*, Birmingham, UK.

Jurafsky, D. (2002): Pragmatics and Computational Linguistics. In *Handbook of Pragmatics*. Blackwell, Oxford.

Kawahara, T. and Hasegawa, M. (2002): Automatic indexing lecture speech by extracting topic-independent discourse markers. In Proc. IEEE-ICASSP, pp. 1–4.

Samy, Doaa, Moreno-Sandoval, Antonio, Guirao, José M. and Alfonseca, Enrique (2006):Building a Multilingual Parallel Corpus Arabic-Spanish-English. In *Proceedings of International Conference on Language Resources and Evaluation LREC-06*, Genoa, Italy

SARIG, L. (1995): Discourse markers in contemporary Arabic. In Zeitschrift für arabische Linguistik , n30, pp. 7-21.

Sperber, D. y Wilson, D. (1986): Relevance: Communication and Cognition. Cambridge. Harvard, University Press

Strassel, S. and Walker, C.R. (2004): Linguistic Resources For Metadata Extraction. In http://www.sainc.com/richtrans2004/uploads/tuesday/Linguistic%20Resources%20for%20Metadata%20Extraction.pdf

Zufferey, S. and Popescu-Belis, A. (2004): Towards automatic identification of discourse markers in dialogs: the case of like. In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, pages 63–71.

Appendix A

Snapshot of the Spanish-Arabic corpus with highlighted discourse markers

1 El Secretario General Adjunto se refirió **en particular** a los progresos logrados en relación con la iniciativa de paz de Djibouti, **así como** a la situación política y humanitaria en Somalia .	1 وأشار وكيل الأمين العام ، التقدم إلى بوجه خاص، الذي أحرز بشأن مبادرة جيبوتي للسلام، بالإضافة الحالة السياسية إلى والإنسانية في الصومال .
2 **Con respecto al** plan de paz de Djibouti , informó a los miembros del Consejo de las actividades del Presidente de Djibouti , Omar Guelleh , en los países de la región ; indicó **también** que la iniciativa había sido bien recibida por la sociedad somalí .	2 وقد أبلغ أعضاء المجلس عن الخطوات التي اتخذت جيبوتي من جانب رئيس في بلدان المنطقة فيما يتعلق بخطة جيبوتي للسلام.
2 **Con respecto al** plan de paz de Djibouti , informó a los miembros del Consejo de las actividades del Presidente de Djibouti , Omar Guelleh , en los países de la región ; indicó **también** que la iniciativa había sido bien recibida por la sociedad somalí .	3 المبادرة أن أيضا وأوضح استقبلت استقبالا طيبا من قبل المجتمع الصومالي.
3 **En cuanto a** la situación política y militar , dijo que durante enero y febrero se habían comunicado incidentes de bandidismo , **así como** confrontaciones entre los clanes .	4 أما فيما يتعلق بالحالة فقد والعسكرية، السياسية ذكر أنه قد ترددت أنباء حوادث لقطع عن وقوع الطرق، بالإضافة إلى العشائر مواجهات بين خلال شهري كانون الثاني يناير و شباط فبراير .