JEP-TALN 2004 - session on Arabic Language Processing
Arabic-English Machine Translation Systems: Discrepancies and Implications

Dr. Sameh Al-Ansary (1) and Prof. Seham El-Kareh (2)

(1) Alexandria University
Phonetics department,
Faculty of Arts,
Alexandria University,
Al-Shatby,
Alexandria,
Egypt.
s.alansary@link.net

(2) Alexandria University,
Phonetics department,
Faculty of Arts,
Alexandria,
Egypt.
selkareh@yahoo.fr

Abstract

This paper will analyze certain translation software and focus on certain linguistic phenomena that are often claimed to be the most problematic areas in machine translation. It will be shown that implementing the system with an adequate parser will empower the translation software. The adopted parsing tool, concentrates on 4 important issues: Constituent vs sentence recognition, Coordination, Preposition semantic function and Prepositional Phrase (PP) attachment. The parser presented is based on a sample corpus of Modern Standard Arabic and a formal grammar to analyze MSA structures automatically. The formal description is implemented using the Affix Grammar over Finite Lattices (AGFL) representing a linguistic approach in terms of functions and categories to account for the sequences inside the structures together with the relations governing these sequences; accuracy exceeded 95%. It is necessary at this point to make two important issues clear. First, in the present stage of the work will be limited to Noun Phrase structures. Second, the adequacy of the grammar is limited to the extent covered by the sample corpus under study.

Keywords

Arabic and computers, Arabic Machine Translation, Arabic Language Processing, Parsing.

1 Introduction

Machine translation research has often been criticized for ignoring developments in linguistic theory. There would appear to be wide communication gab between theoretical linguistics and practical machine translation research. Some observers believe that there are good reasons for this situation: until recently linguistic theories had not provided adequate accounts of all aspects of language use; a good linguistic theory may have given a convincing analysis of, say, quantifiers or coordination but not explained all the peculiarities of actual usage in the coverage required for machine translation. However, recent theories such as Lexical Functional Grammar (Shihadah et al 1998). or Generalized Phrase Structure Grammar (Gazdar et al 1985) and their various derivatives have set out explicitly to cover as broad a range as possible, not only within one specific language but also for different types of languages. In the past, and unfortunately, it is generally true today, much of linguistic theory was based on phenomena observed in English, the language of the majority of theoretical linguistics. This neglect of other languages has been why linguistic theory has had less impact on Machine Translation than some observers might have expected. In other words, linguistic theories have rarely addressed questions of contrastive linguistics, i.e. the way in which different languages use different means to express similar meanings and intentions. Such questions are of course at the heart of Machine Translation.

2 Linguistic and Formal Framework

In this section the linguistic and formal framework is going to be explained briefly in which a NP category occurred which represents the core of our parsing tool that can be used efficiently in translating Arabic structures into English.

Analyzing linguistically the NP, a number of functions could be distinguished. These functions are: the head, the determiners and the postmodifier. The individual behavior of these functions ranges between determination and postmodification of the nucleus of the NP, the element occurs in the head function. The NP, in its simplest form , only consists of a head. Even in a more or less complicated NP, only one head function is to be distinguished. The head is specified and defined as the unit which is marked for its function at the next higher level of description and cannot be deleted without affecting the meaning of the constituent (Al-Ansary 2002), (Ditters 1992)). By this definition, the head function of an NP can only be realized by the category noun. Owens (1988) distinguished several subcategories able to realize this function (cf. Owen's 1984, Ch. 9). According to our subclassification of nouns, a common noun, pronoun, proper noun, present participle, passive participle, adjectival noun, standard infinitive (verbal noun), noun of title… etc are examples of heads of an NP (for more details cf. (Al-Ansary 2002) and (El-Kareh S., Al-Ansary S. 2000). In extension of the head, an element can function as a determiner to the head of the NP. The element occupying this function may occur before or after the head. This brings us to differentiate between what is called a "predeterminer" (PREDET) and "postdeterminer" (POD). However, it has to be kept in mind that they are mutually exclusive in relation to the head i.e. they could not occur together. The category in the function of predeterminer is mainly the prefixed article "الـ" while the category in the function of post-determiner is a normal NP marked for genitive case. The post-modifier function is always placed after the head of the NP and; is for this reason called "post modifier" (POM). In our approach, post-modification could, according to its categorial realization, be classified into PPOM, ADJPOM, NPOM or ADVPOM, realized by a

prepositional phrase, adjective phrase, noun phrase and adverbial phrase respectively. An additional element could be distinguished functioning as a complement of the head of the NP (COMPL). Like post-modification, the complement function is always realized after the head. However, it is not recommended to treat both of them as a post-modification since the complement has a particular syntactic function in relation to the head. For example, a post-modifier follows its head with respect to 'definiteness', 'number', 'gender', and 'case'. There is no direct relation between the head and its complement as far as agreement is concerned. On the contrary, the head imposes specific values on its complement.

Formally, the linguistic description of the NP in MSA can be represented by means of context-free rules. To implement the formal grammar, a two level approach for syntactic description was used by means of the AGFL (Affix Grammar over Finite Lattices) formalism. In this way, ROOT is the start symbol in our grammar and ROOT is rewritten in the phrasal category NP. The NP is in its turn rewritten as a sequence of optional and obligatory functional elements (which constitutes the first level of description, syntactic level). The description alternates between functions and categories till the description in lexical terms has been reached. Starting with our initial label ROOT, the first rule is: ROOT: NP. A number of restrictions is applied via some linguistic features to determine the dependencies and relations between the elements of the NP (which constitutes the second level of description, affix level). Thus our start label could be revised as: ROOT: NP(definiteness, number,gender,person,case). By means of the non-terminal affix variables, the elements of the first and second levels of description are dealt with. At the first level, non-terminal elements are arranged in phrase structure rules called syntax rules or hyper rules. These phrase structure rules are context-free rules describing syntactic structures. As it has been seen with 'definiteness', number, gender, person and case, other affix variables can be attached to the non-terminal of the first level. These meta affixes constitute the second level of description. For more details about conventions for writing rules of AGFL confer Koster (1991).

3 Important Issues

In this section some linguistic phenomena that highlight important problematic issues in the field of Arabic-English Machine Translation will be presented. These linguistic phenomena are : Constituent vs sentence recognition, Coordination, Preposition semantic function and PP attachment. It will be demonstrated how weak the existing translation systems are and how the parsing accuracy of the parser adopted can contribute to reach an acceptable translation.

3.1 Constituent vs Sentence Recognition

One of the important issues that a Machine Translation system should take care of is the ability to differentiate between a phrase and a sentence. In Arabic, a nominal sentence can be composed of a topic and a comment. The comment can be realized by a PP, giving a complete meaning of the whole structure, the sentence. Consider the example in (1):

(1) محمد في المدرسة moħammadun fi ?almadrasati 'Mohammad is at school'

In this example, "محمد" (NP) is the topic of the sentence while "في المدرسة" (PP) is the comment of the sentence. Thus the translation given above can be considered as an acceptable translation of this sentence. This kind of sentence structure can cause a formal ambiguity

contradicting with phrase structures. Consider examples in (2a) and (3a) with their automatic translation (Program 1 is InterNet Translation Service from CIMOS Company. http://www.cimos.com/. Program 2 is Ajeeb: http://tarjim.ajeeb.com/ajeeb/) :

(2) a) القبض على ٥ إرهابيين ?alqabdu ʕala xamsati ?irha:bijj:n
Program 1:* The arrest is on terrorists Program 2:* taking possession of 5 terrorists

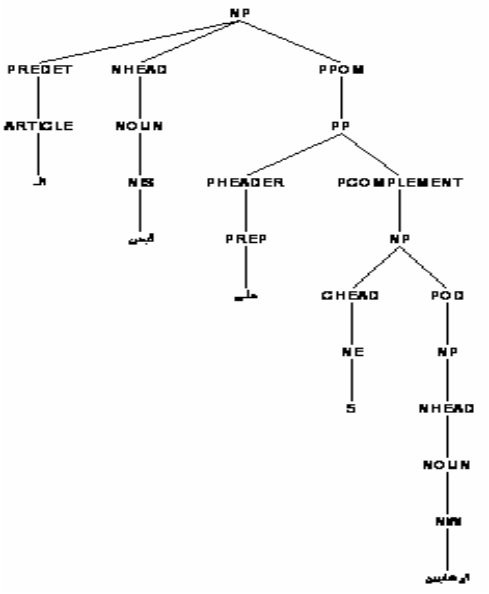
b) The arrest of 5 terrorists

(3) a) تمكن الفكر المتطرف من أصحابه tamakkunu ?alfikri ?almotaṭarrifi min ?asħa:bihi
Program 1: * A mastery the extreme thinking is from his owners
Program 2:* The extreme thinking , take possession of his companions

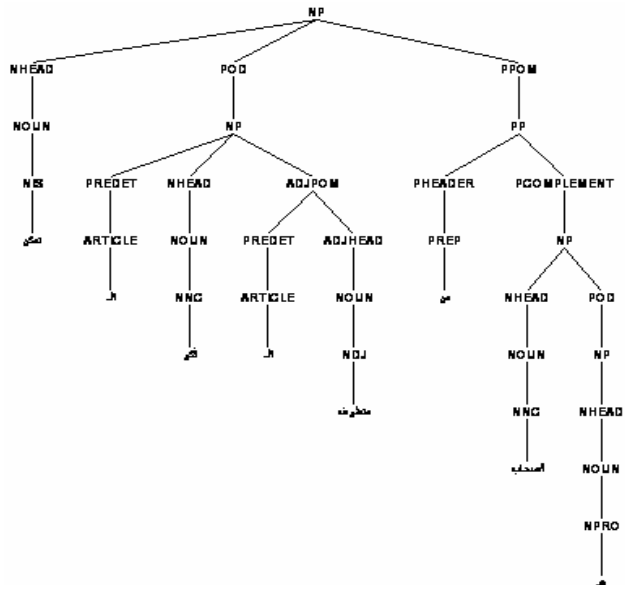
b) Extremist thinking has dominated them

Each of (1) , (2a) and (3a) consists of NP + PP, however the structure in (1) is a sentence while the structures in (2a) and (3a) are NPs. The false translation is caused by the software in which the software is unable to recognize and differentiate between constituent structure and sentence structure. The contribution of our parsing tool resides in using our linguistic and formal strategies to give a reliable representation. This could be examined through the labelled tree representation in (4), consequently a reliable translation as those given in (2b) and (3b) can be obtained.

(4) (a)



(4) (b)



3.2 Coordination

Coordination is a problematic issue in almost all languages. It has been assume that coordination takes place at the categorical level and that it is dominated by a single function node. It has also been assumed that the coordination concerns similar categories. In the present stage of the formal description of coordination, the work has been limited to detect the boundaries of the coordinated NPs without going into details of revealing all information that

can result from coordinating a NP with another and how they can affect the identity of the whole constituent (In writing a formal grammar for describing coordination, the grammar should consider the result of coordinating the features of the first NP with those of the second NP. For example, considering definiteness when the first NP is definite and the second is indefinite, the coordination will result in a definite constituent. Thus, the formal description of coordination should deal with features like definiteness, number, gender, and person to reveal how they can affect the whole constituent. For details about this kind of investigations cf (Ditters 1992).) To enable our description to detect automatically the boundaries of two coordinated NPs, all available information at hand have been made used of. Since the coordinator separates two NPs, the grammar tried to get all possible alternative combinations that can be expressed before and after the coordinator. Thus a more accurate description to the NP that precedes the conjunction has been needed to enables precise division in one hand and eliminate the rest of the alternatives on the other. Up to the NPs analyzed the affixes 'definiteness', 'subclass' and 'case' have been used to control the limits of the NPs being coordinated. In (5a) a structure that has a coordination at a certain linguistic level i.e. a postdetermination level of the Nominal Head 'مجلسي' can be seen. A full parsing in labelled tree of this structure is presented in (6). This structure has been tried to be translated automatically, the result can be seen in (5a).

(5) (a) عدد من أعضاء مجلسي الشيوخ والنواب

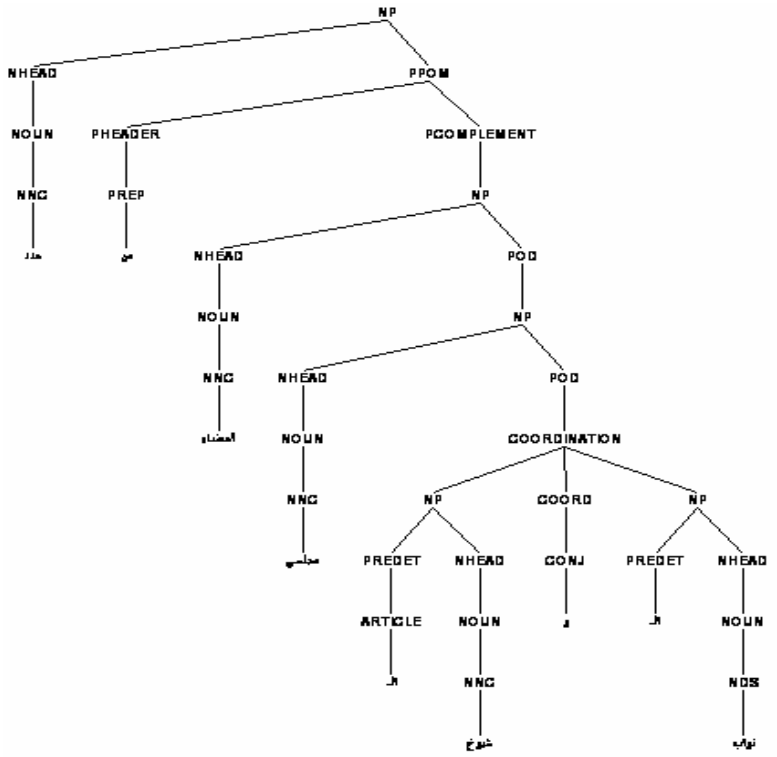
ʕadadun min ʔaʕḏa:ʔi maglisajj ʔašiju:xi wannowwa:bi

Program 1: * Number of the sheikhs councils members and the deputies.

Program2 : * He wailed from the councilors of the whitebeards and the vicegerents

(b) A number of the members of sheikhs and the deputy councils.

(6)



It is very clear from the acceptable translation in (5b) that the boundaries of the coordinated NPs are not precise. The following brackets in (7) can show the boundaries of the coordinated parts according to that translation.

(7) عدد من[[أعضاء مجلسي الشيوخ] و[النواب]]

However, the parsing tool adopted, relying on subclass, definiteness, case of the coordinates, and the number of the Nominal Head 'maglisajj' to which the coordinates are postdetermining, could reliably detect the coordinated parts as shown in (8).

(8) عدد من أعضاء مجلسي [[الشيوخ] و[النواب]]

Consequently, relaying on the parsing tool adopted a reliable translation can be obtained as that given in (5b).

3.3 Preposition Semantic Function

It is very important for a good translation system to automatically detect the meaning of a preposition in a structure. This issue is very dangerous in the translation from and to Arabic because the same preposition in Arabic can convey more than one semantic meaning. Consider the following examples:

(9) (a) رسالتان من مبارك لزايد وجابر الأحمد risa:lata:n min muba:rak liza:jed waga:bir ?al?ahmad
'Two letters from Mubarak to zayed and gaber Alahmed'

(b) أعلى نسبة مشاهدة لحوار تليفزيوني في العالم
?aʕla: nisbati muša:hadatin lihiwa:rin tilifizjo:nijjin fi ?alʕa:lami
'A highest ratio of viewing for a TV talk in the world'

In (9a) the preposition "لـ" prefixed to "زايد" has a target meaning since the structure has a source "من" preceding. While in (9b) the same preposition has an associative meaning connected to the superlative noun, the Head of the NP. However, it has been noticed that Arabic Machine translation systems have fixed the semantic function of prepositions. Consider the following examples with their corresponding translation

(10) (a) قادة الجبهة الإسلامية للإنقاذ qa:datu ?algabhati ?al?isla:miyyati lil?inqa:ði
Program 1: * The leaders of the Islam bloc to the Rescue
Program 2 : * The commanders of the Islamic front of the deliverance

(b) محاولة لتوجيه المتحدث إلى إدانة هذا الفكر وهذه الجرائم
muħa:walatun litawgi:hi ?almutaħaddiθi ?ila ?ida:nati ha:θa ?alfikri waha:ðihi ?algara:?imi

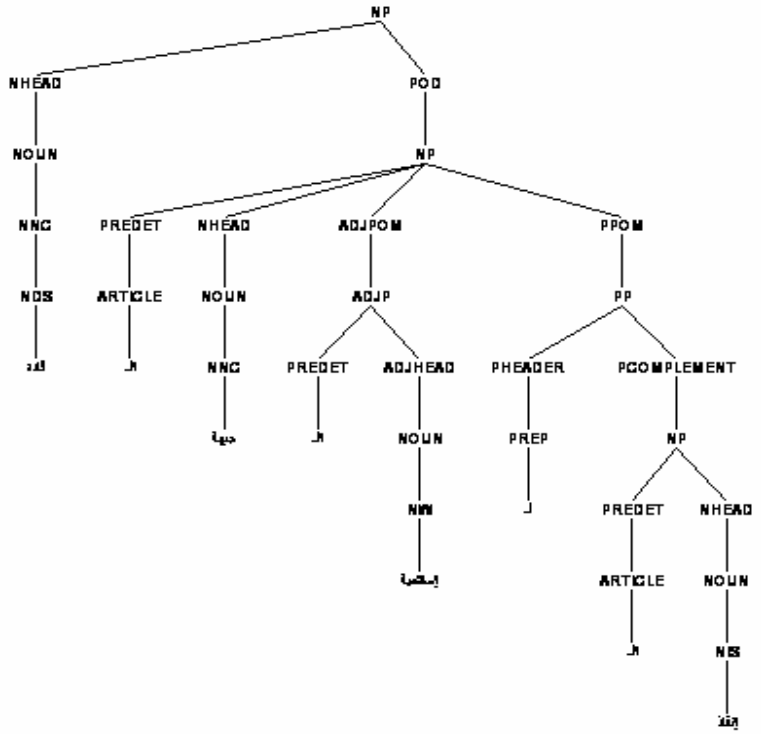
Program 1: *An attempt to the speaker directing to the conviction of this thinking and these crimes

Program 2: * An attempt to crown him the spokesman to condemnation of this thinking and this the crimes.

(c) زيارة لأمريكا zija:ratun li?amirika
Program 1: A visit to America. Program 2: A visit to America.

As it is clear from the examples listed in (10) together with their automated translation that the meaning of the preposition "لـ" is always fixed, thus it was incorrect in (10a, b) and acceptable, accidentally in (10c). The parsing system adopted could link the meaning of the preposition with the structure it occurred in leading to a reliable translation in this respect. In (11) an example of a labelled tree representation of (10a) can be seen:

(11)



In this analysis the parser could detect the associative meaning[1] of the preposition (لـ) which directly leads to translate is as "for" not "to".

3.4 PP Attachment

Attaching the PP to a given part in the structure can make a kind of ambiguity that affects translating that structure from Arabic to English. Consider the structure in (12). In fact if you concentrate for a while, you will notice that it can support two translations depending on the attachment of the PP (في الشرق الأوسط). This PP can be considered as a locative constituent attached to عقد or as a prepositional postmodifier attaching this constituent to سلام. In fact parsing tool can give flexiblely two parsing trees that can be implemented in two different translations. The discourse will the exact translation. In (13a) and (14a) two parse trees (in labelled brackets format) were given with their translations in (13b) and (14b) according to the structure.

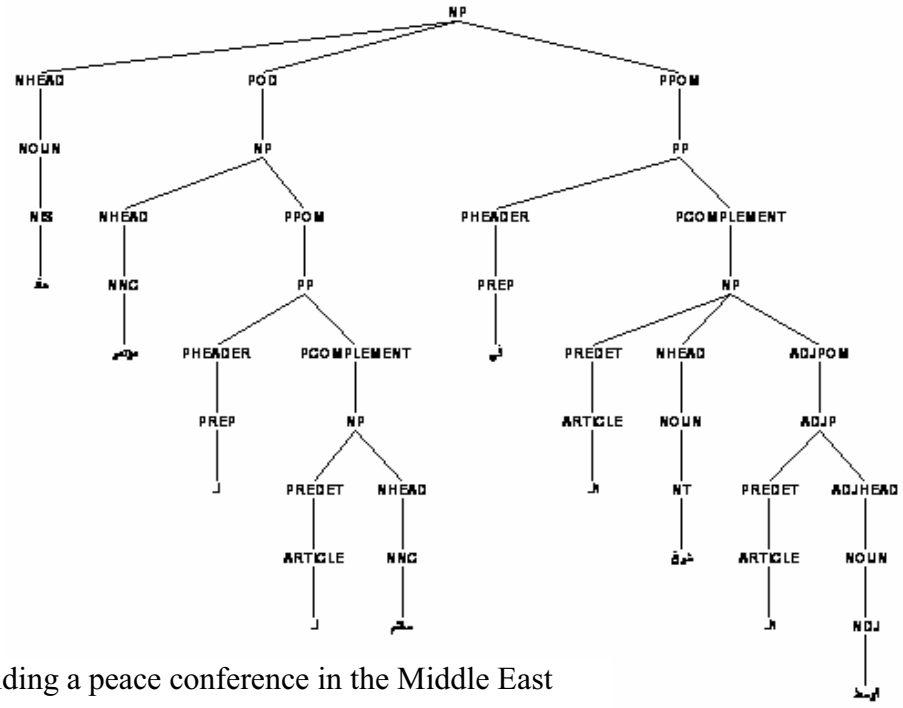
---

[1] In the tree diagram above, it is not clear how the parser could discover the semantic meaning of the preposition. The output of the parser in labelled bracketing format is as follows in which the semantic function of the preposition is clear: NP(NHEAD(NOUN(NDS(قادة))),POD(NP(PREDET(ARTICLE(الـ)),NHEAD(NOUN(NNC(جبهة))),ADJPOM(ADJP(PREDET (ARTICLE(الـ)),ADJHEAD(NOUN(NW(إسلامية))))),PPOM(PP(PHEADER(**PREP(ASSOCIATIVE**لـ)),PCOMPLEMENT (NP(PREDET(ARTICLE(لـ)),NHEAD(NOUN(NIS(إنقاذ)))))))))

(12) عقد مؤتمر للسلام في الشرق الأوسط

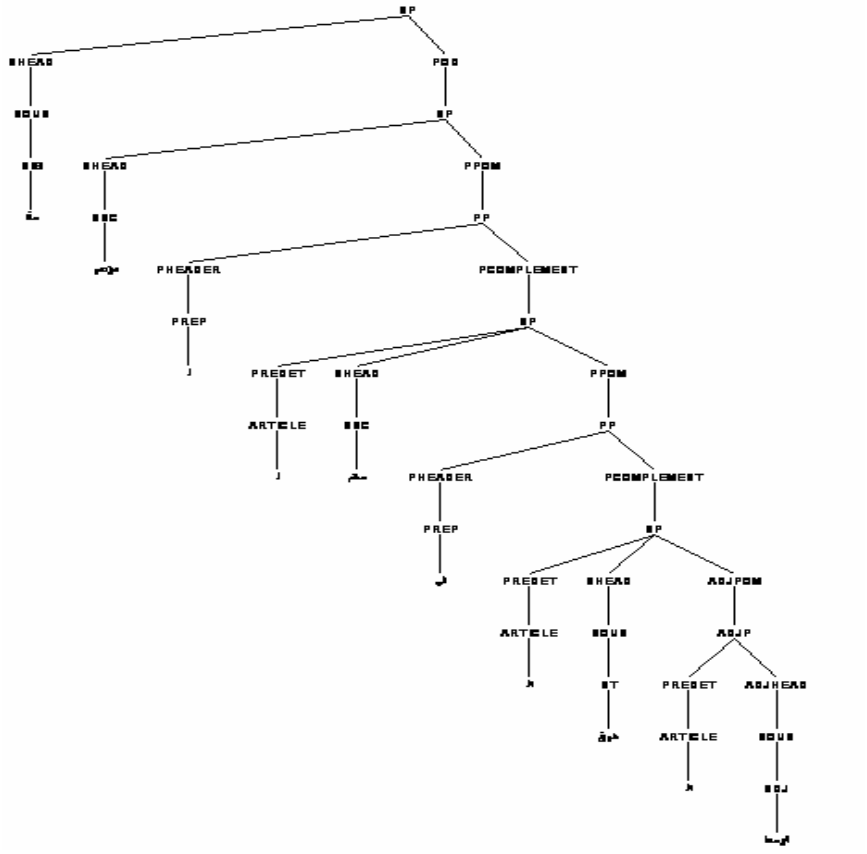
ʕaqdu mo?tamarin lilsala:mi fi ?alšarqi ?al?aws̱at̠i

(13)(a)



(b) Holding a peace conference in the Middle East

(14) (a)



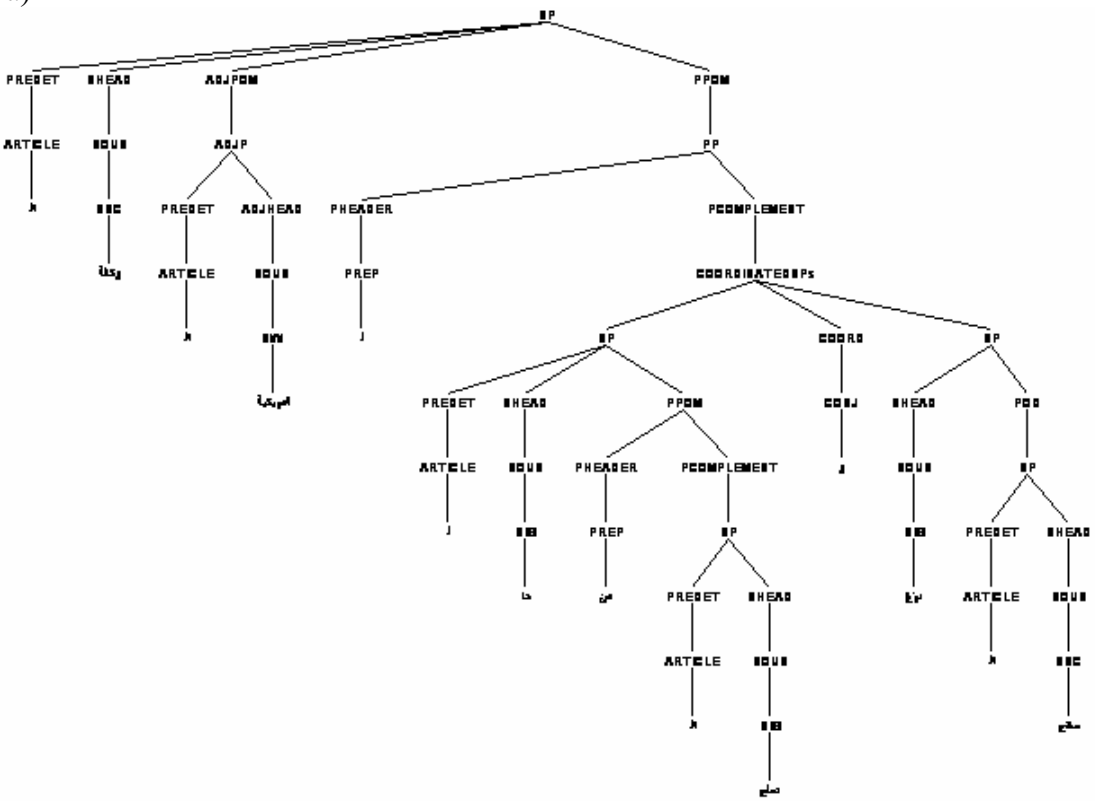
b) Holding a conference in the Middle East about peace

The problem becomes more severe when the prepositional complement is realized by a coordinated NP. In (15) we see the nominal head "وكالة" of the NP is postmodified by an ADJP "الأمريكية" then by a PP. According to our linguistic approach this PP is composed of PHEADER and PCOMPLEMENT functions. The latter is realized by a coordinated NP. A full labelled brackets representation describing the whole structure can be seen in (16a). This structure is tested over a famous Arabic-English translation system, the result was as represented in (16b).

(15) الوكالة الأمريكية للحد من التسلح ونزع السلاح

?alwika:latu ?al?amri:kijjatu lilħaddi mina ?altasaluħi wanazʕi ?alsila:ħi

(16) a)



b) Program 1: *The American agency to the restriction is from the arming and the weapon pulling out

Program 2: *The American stewardship to be anatheist from the armament and the demilitarisation

c) The American agency for limiting arms and disarming.

The labelled tree representation above links the NHEAD with its modifiers in a syntactic harmony that can lead to an acceptable semantic interpretation of the whole structure as it appear in the translation in (16c).

3.5 Conclusion

This paper has focused on Machine Translation of Arabic in which the Arabic language will be a source language. What the paper tried to prove is that in order to achieve a good machine

translation for Arabic, in principle, Arabic structures should be understood by the machine. Using the parser adopted, it was possible to focus on building a formal module for describing Arabic structures and transfer them with the same meaning into the target language. Some problematic areas in Arabic-English machine translation have been surveyed. To a certain extent, the gabs in the translations have been showed, and how the parsing system adopted could deal with these gabs. Work in this direction will help to form a new generation of Arabic studies using Information and Communication Technology for research and teaching purposes.

References

Al-Ansary S. (2002), A Comparative Corpus-based Study of Spoken and Written Modern Standard Arabic (MSA), Ph. D thesis, Alexandria University, Egypt.

Ditters W. E. (1992), *A Formal Approach to Arabic syntax: The Noun Phrase and Verb Phrase*. Ph.D, Nijmegen University.

El-Kareh S., Al-Ansary S. (2000), An Interactive Multi-Features POS Tagger. In *the Proceedings of the International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Intelligence for Decision Control and Automation in Engineering and Industrial Applications,* Natural Language Processing Panel, pp. 83-88. 22-24 March 2000, Monastir, Tunisia.

McEnery Tony, Andrew Wilson (1993), *Corpora and Translation: uses and future prospects,* Lancaster: UCREL

Nirenburg, Sergei et al. (1992). *Machine Translation: A Knowledge-based Approach*. San Mateo, Cal.: Morgan Kaufmann.

Newton, John (ed.) (1992). *Computers in Translation: A practical Appraisal*. London : Routledge

Owens, Jonathan (1988), *The Foundations of Grammar: An Introduction to Medieval Arabic Grammatical Theory,* John Benjamin publication company, Amsterdam.

Shihadah M., Paul Roohnik (1998), Lexical Functional Grammar as a Computational-Linguistic Underpinning to Arabic Machine Translation. *The proceedings of the 6th International Conference and Exhibition on Mutilingual Computing*, University of Cambridge, London, 17-18 April 1998, pp. 5.8.1 – 5.8.9

Schubert, Klaus (1987), *Contrastive Dependency Syntax for Machine Translation,* Dordrecht: Foris Publications.

Sigurd, Bengt (1994), *Computerized Grammars for Analysis and Machine Translation*. Lund : Lund University Press.

Trujillo, Arturo (1999), *Translation Engines: Techniques for Machine Translation,* London.

Whitelock, Peter and kieran Kilby (1995), *Linguistic and Computational Techniques in Machine Translation System Design*. London: UCL Press.