Processing Arabic Text Content: The Encoding Component in an Interlingual System for Man-Machine Communication in Natural Language[1].

Sameh Alansary[†][‡] Sameh.alansary@bibalex.org

Magdy Nagi[††][‡] magdy.nagi@bibalex.org

Noha Adly[††][‡] noha.adly@bibalex.org

[‡] Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.

[†] Department of Phonetics and Linguistics
Faculty of Arts
Alexandria University
El Shatby, Alexandria, Egypt.

[††] Computer and System Engineering Dept.
Faculty of Engineering
Alexandria University,
Egypt.

1. Abstract.

It could be true that any attempt to build a system to process the content of a given text written in a given language will be faced by tackling language analysis tasks. To reach the semantic representation of any sentence, the system should be enriched with a technique for lexical and syntactic disambiguation. Having finished with semantic representation, the system should be able to re-synthesize the semantic representation into another acceptable sentence in the target language. However, it is not that easy; there are many problems that need to be solved in both the analysis and synthesis processes.

To avoid the pitfalls associated with approaches relying on intermediate representations, e.g. syntactic tree, this paper presents an approach on which processing Arabic content, and even the exchange of information among languages, starts directly from a semantic layer without passing through the level of syntax. The approach encodes Arabic structures into a set of semantic relations between a set of nodes representing the elements (words) of the sentence as concepts. Once the concepts are built, the relations between them are determined and can be decoded again to any other language.

The grammar for the encoding process is implemented in Universal Networking Language (UNL); it enables computers to understand natural languages which will make it possible for humans to communicate with machines in natural language. Encoding Arabic sentences in terms of semantic networks depends mainly on holding theta roles between different arguments (the participants of the event or situation) included in the sentence. Therefore, the arguments of the predicates of the natural language are classified into a closed set of types which have a different status in the grammar.

2. Introduction.

The need for efficient and instantaneous information exchange across languages is growing by the day. It has been hoped that this exchange of information can occur in the native language of the user. Also, most of the current documents, representing scientific and educational information, are written in English or in a number of other languages. This makes the benefits of these documents limited to the native speakers of these languages; non-native speakers of these languages have to learn them to be able to use these documents. However, it would be great if everyone can access these documents in everyone's mother tongue.

Many distinct approaches have been used to automate processing content between NLs. Perhaps, the most obvious approach is to use an intermediate representation, e.g. a syntax tree. For example, to process English in German could start with an English grammar to produce syntax trees. Inputting an English sentence using an English lexicon should produce a syntax tree for the English input (for more detailed problems of syntactically based machine translation systems cf. Al-Ansary and El-Kareh (2004a)). The 'base' words (lexemes) in the tree could then be mapped to their German equivalents, and the resulting syntax tree for German could be put back through the grammar (using a German lexicon) to yield a German sentence (Hutchins et al (1992), Eynde (1993), Arnold (1994) and Nirenburg et al (2003)). For some pairs of sentences this kind of approach clearly works. Note that to translate among a set of languages we need a 'translation lexicon' for each pair, so for 10 languages we would need 90 such lexicons. In addition, we would need different grammars, since word orders differ among languages. Attractive though it seems, the whole approach is

[1] This work was carried out within Ibrahim Shihata Arabic-UNL Center (ISAUC) hosted by Bibliotheca Alexandrina (BA) and managed by its affiliated research center, the International School of Information Science (ISIS). The project is funded by the Arab Fund for Economic and Social Development, Kuwait. ISAUC is playing a major role in designing and implementing the Arabic language tools to act as an active language center for Arabic.

fundamentally defective because of two reasons. First, it is not possible to translate lexemes 1:1 between languages because of lexical ambiguities. In every language, words have ranges of meaning, but the ranges differ. For example, in English, 'wood' has two quite distinct meanings: the substance of which trees are composed, and a piece of land on which trees grow. Translation into French can (apparently) be 1:1 because the French word 'bois' has the same two meanings. But German has two distinct words: 'Holz' for the substance and 'Wald' for the piece of land. Hence, the sentence: 'The man sees the wood' can be translated into French by the method described above, but not into German. As a minimum, then, contextual information from the sentence and passage being translated will be needed to determine the correct translation of the lexemes. However, even this does not work. Consider the sentence 'That's not a wood, that's a forest'. This can be translated into French as 'Ce n'est pas un bois, c'est un forêt', since French, like English, has words for small and large pieces of land covered with trees. However, German uses only the single word Wald. So how could the English and French sentences be translated 1:1 into German? Even with contextual information, 1:1 word translation is not possible. Second, different languages, even within the same language family, use different syntactic structures to convey the 'same' meaning. For example, the normal way of saying 'I like dancing ' in Spanish or Modern Greek demands the use of a totally different structure. The literal translation of the Spanish 'me gusta bailar' is 'me it-pleases to-dance'; the literal translation of the Greek 'mou aresi na horevo' is 'of-me it-pleases to I-dance'. It is possible to imagine the method described above translating between the Spanish and Greek sentences, but not between English and either of the other two languages (Dorr (1993)).

Each language needs its own grammar and lexicon that generate its own syntax tree (or other intermediate representation). Each pair of languages then needs its own translation lexicon (which will need input from semantic-based processing of the sentence and the passage to determine context), and its own translation rules (which will specify appropriate changes of syntax for some semantically 'equivalent' sentences). But for 10 languages we now have 10 Grammars, 10 Lexicons, 90 translation lexicons and 90 sets of translation rules, plus the need for semantic processing in each language.

The situation will be completely different with the UNL. The UNL is a formalism for computers that is used to represent sentences considering their logical relations without lexical and syntactic ambiguities (see section 3). It has been designed to intermediate between natural languages allowing people with different native languages to communicate with each other over the net every one in his own native language (Uchida (2001)). Therefore, it could be very powerful if it can be used to process the content of any text whatever the language of the text is because it is free of lexical ambiguities as it works with "Universal Words"; this represents universal concepts regardless of any language. If it exists in a given Natural Language that a lexical item has more than one meaning, each meaning will be represented in UNL as a completely different Universal Word. This means that we can translate 1:1 lexeme in two languages safely if both lexemes have the same 'Universal Word'. In this concern UNL can be considered economic in the sense that only 10 translation dictionaries will be needed to translate among 10 languages, instead of the 90 translation lexicons needed in the other approach used by other systems. This way, UNL can break language barriers and contribute in minimizing the digital divide (Auh (2001), Galinski (2001)), building infrastructure for human-machine communication in natural language (Hausser (1999, 2001)), and building information societies (Montviloff (2001)). Therefore, we can live all together in the cyberspace communicating with each other through machines in our native languages.

The Bibliotheca Alexandrina (BA) is dedicated to recapture the spirit of the ancient Library of Alexandria, a center of excellence for world learning. The Library and its affiliated research center, the International School of Information Science (ISIS), are devoted to using the newest technology to build a universal digital library accessible to people worldwide. ISIS has worked on a number of projects such as the Million Book Project Saleh et al (2005), digitizing over 27,000 books and publishing them in searchable form online. Other projects include the Nasser Digital Library Eldakar et al (2005), which contains thousands of digital resources in multiple formats, the Dar Al-Hilal Digitization project, Description de l'Egypte, and other projects related to the Digital Library of the Modern History of Egypt. Moreover, ISIS has been working on integrating all the Library's digital assets into a huge Digital Assets Repository (DAR)[2], providing public access to digitized collections through web-based search and browsing facilities, and has developed the Universal Digital Book Encoder (UDBE) Eldakar et al (2006), facilitating the electronic publishing of multi-lingual electronic documents.

The UNL system will play a major role in the dissemination of knowledge and offer a powerful platform for inter-lingual communication in furtherance of the purpose of sustainable development, intercultural understanding and dialogue, and peaceful sharing of economic and social world resources. ISAUC will augment the vision of the BA in becoming an international center of excellence using the latest technology to make the works of man accessible to people worldwide. The success of UNL will increase the propagation of BA's digital library through the conversion of Arabic digital resources such as the Nasser collection to multiple languages through the Internet, in addition to the universal use of the BA Library Information System.

[2] The Digital Archive of President Gamal Abdel-Nasser. Available at : http://nasser.bibalex.org

This paper describes the encoding module in a man-machine communication system that encodes Arabic text input to a hyper semantic network according to the Universal Networking Language (UNL) framework. It is organized as follows: Section 3 presents the Universal Networking Language. Section 4 describes the structure of the Arabic dictionary used in the 'Encoding process' (or 'EnConversion process' in UNL technical terms). Section 5 is the main section that discusses how Arabic structures are EnConverted to UNL expressions. Section 6 presents a conclusion and future work. The decoding component that decodes the resultant UNL expression to natural language is dealt with in Alansary et al (2006) in this volume.

3. The UNL System.

UNL is an acronym for "Universal Networking Language". It is an artificial language in the form of semantic network for computers that resides in the cyber world to express and exchange every kind of information. UNL is a "language" for computers (different from a "computer language") that expresses information and knowledge in digits, the characters that all computers understand. It is capable of encoding contents from any natural language into UNL and then decodes it into any other natural language. UNL Language enables peoples to build the "banks" of human knowledge from and to diverse natural languages (Uchida (1996), Uchida et al (2002a), Uchida et al (2005)).

UNL represents information, i.e. meaning, sentence by sentence. Sentence information is represented as a hyper-graph having Universal Words (UWs) (See section 3.2.1) as nodes and relations as arcs. This hyper-graph is also represented as a set of directed binary relations, each between two of the UWs presented in the sentence. In addition to the core meaning of the sentence, the UNL expresses information classifying objectivity and subjectivity. Objectivity is expressed using UWs and relations. Subjectivity is expressed using attributes by attaching them to UWs. The following subsections go in some more details with the UNL system to an extent to make it easy to deal with detailed formal description of Arabic in UNL formalism.

3.1 The UNL System Components

The UNL System consists of three major components:

1) Language Resources: They are divided into language dependent part and language independent part. Linguistic knowledge on concepts that is universal to every language is considered as language independent and to be stored in the common database, UNLKB. Language dependent resources like word dictionaries and rules, as well as the software for language processing, are stored in each language server. Language servers are connected through the Internet. Supporting tools for producing UNL documents can be used in a local PC. Such supporting tools operate with consulting language servers through the Internet. Verification of UNL documents can be carried out through the Internet or in a local PC. UW Gate for searching and maintaining the common database UNLKB operates through the Internet. Figure (1) shows the structure of the UNL system (For more details about the structure of the system, cf. Uchida (1996) and Uchida (2002b)).

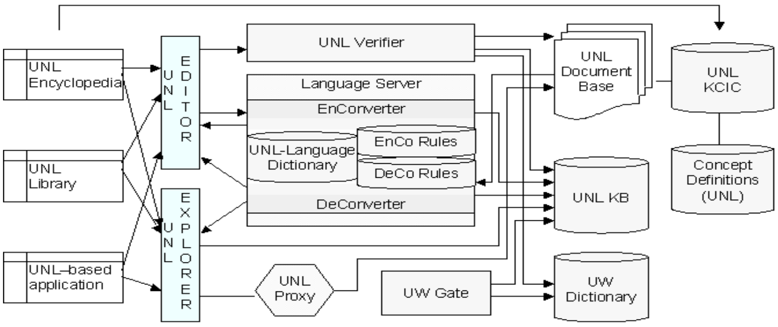


Figure (1): Structure of the UNL System

2) UNL Converters: They are engines used to develop a grammar for each language to store it on each language server which are connected through the Internet. These converters are the **EnConverter** and **DeConverter** which are the core software in the UNL system. The EnConverter converts natural language sentences into UNL expressions. The Universal Parser (UP) is a specialized version of the EnConverter. It generates UNL expressions from annotated sentences using the UW dictionary without using grammatical features. All UNL expressions are verified by the UNL verifier. The DeConverter converts UNL expressions to natural language sentences. Figure (2) shows the mechanism how a UNL document is made and how a UNL document is converted into natural languages in the UNL system. Arrows in solid line show dataflow, arrows in broken line show access.

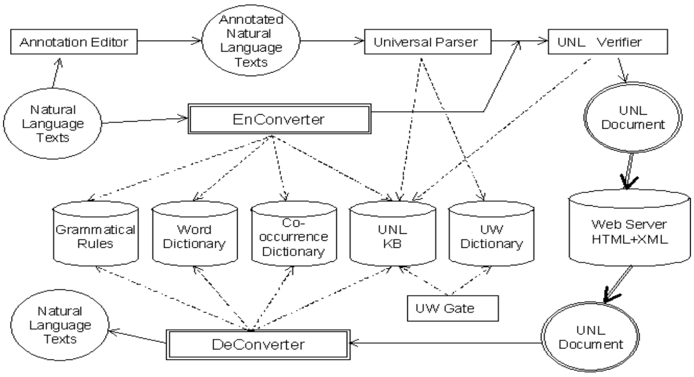


Figure (2): Mechanism of conversion of UNL

a) EnConverter:

Enconverter is a language independent parser that provides synchronously a framework for morphological, syntactic and semantic analysis. It is designed to achieve the task of transferring the natural language to the UNL format or UNL expressions which are semantic networks made up of a set of binary relations, each binary relation is composed of a relation and two UWs that hold the relation. A binary relation of UNL is expressed in the following way:

<relation> (<uw1>, <uw2>)

EnConverter should work for any language by simply adapting a different set of the grammatical rules and Word Dictionary of a language. For this purpose, the function of EnConverter should be powerful enough to deal with a variety of natural languages but never depend on any specific language. As a result, the EnConversion capability of EnConverter covers context-free languages, as well as context-sensitive languages.

EnConverter checks the formats of rules, and outputs messages for any errors. It also outputs the information required for each stage of EnConversion in different levels. With these facilities, a rule developer can easily develop and improve rules by using EnConverter. In general, figure (3) shows how EnConverter works.

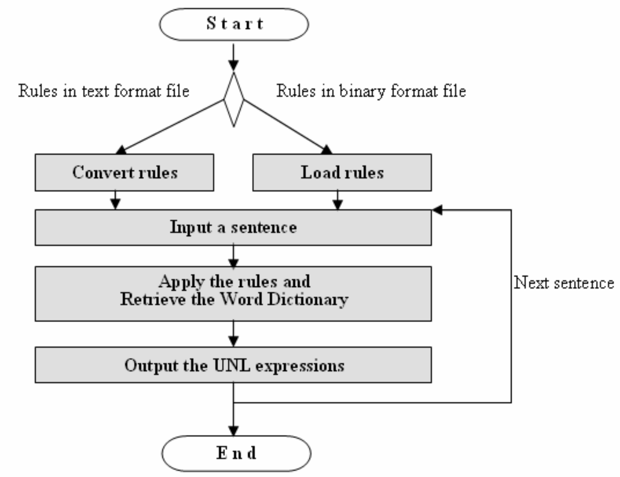

Start
Rules in text format file
Rules in binary format file
Convert rules
Load rules
Input a sentence
Apply the rules and
Retrieve the Word Dictionary
Next sentence
Output the UNL expressions
End

Figure (3): Flowchart of EnConversion process

Firstly, converts EnConversion rules from text format into binary format, or loads the binary format EnConversion rules directly if the rule file is already converted to binary format. Secondly, it inputs a string or a list of morphemes/words of a sentence of a native language. Thirdly, it starts to apply rules to the Node-list from the initial state (the starting node of the node list representing the input sentence, see figure (4)).

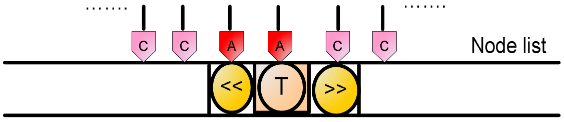


Figure (4): Initial state of the Analysis Windows and the Node-list.

Figure (4) shows the initial state of the Analysis Windows and the Node-list, when the text of a sentence is input. The current Analysis Windows (A) are on the Sentence Head node (<<) and the input text (T). If a list of morphemes is input, the Analysis Windows will be placed on the Sentence Head node (<<) and the extreme left node, namely the first morpheme of the input. EnConverter applies EnConversion rules to the Node-list through its windows. The process of rule application is to find a suitable rule and to take actions on the Node-list in order to create a UNL network using the nodes in the Analysis Windows. It stops when either the Sentence Tail (>>) moves to the left Analysis Window or the Sentence Head moves to the right Analysis Window (See figure (5)). Finally, the UNL network (Node-net) is output to the output file in the binary relation format of UNL expression.

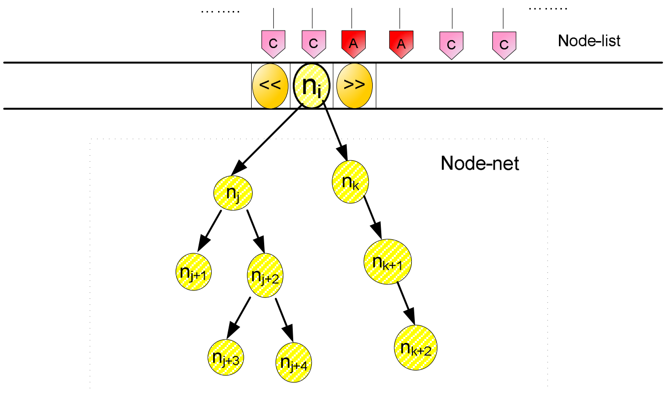


Figure (5): Final state of the Analysis Windows and the Node-net.

b) DeConverter:

DeConverter is a language independent generator that provides synchronously a framework for morphological and syntactic generation, and word selection for natural collocation. DeConverter can deconvert UNL expressions into a variety of native languages, using a different set of files such as the Word Dictionary, Grammatical Rules and Co-occurrence Dictionary of each language. DeConverter generates target sentences of a native language from UNL expressions by applying deconversion rules. (cf. Uchida (1996, 2001) and Alansary et al (2006) of this volume for more information on the Deconverter)

3) Supporting Tools: This is for producing UNL documents. They can be used on a local PC. Such supporting tools operate with consulting language servers through the Internet. Verification of UNL documents can be carried out through the Internet or on a local PC. UW Gate for searching and maintaining the common database UNLKB operates through the Internet.

3.2 UNL Language Components:

The UNL consists of Universal words (UWs), Relations, Attributes, and UNL Knowledge Base. The Universal words constitute the vocabulary of the UNL, Relations and attribute constitutes the syntax of the UNL and UNL Knowledge Base constitutes the semantics of the UNL. The following subsection will deal with UWs, Relations and Attributes respectively. The section will be ended with a concrete example of UNL graph.

3.2.1 Universal Words (UWs): The Vocabulary of UNL.

A Universal Word represents simple or compound concepts. UWs are made up of a character string (an English-language word) followed by a list of constraints. There are three kinds of UWs. Basic UWs, Restricted UWs and Extra UWs (Uchida et al (2002)).

3.2.2 Relations: The Syntax of UNL.

Binary relations are the building blocks of UNL sentences. They are made up of a relation and two UWs. The relations between UWs in binary relations have different labels according to the different roles they play. The relations are linguistically (semantically) based (Uchida (2003)) and similar to those described by Fillmore (1969). A relation label is represented as strings of 3 characters or less, see the following example:

agt (agent) relation: It indicates a thing in focus that initiates an action. An agent is defined as the relation between UW1 (do) and UW2 (a thing) where UW2 initiates UW1. Consider the following sentence: John breaks the glass. As "John" is the initiator of the action (a thing) and "break" is the event (do), then UW1 will be represented by "break" and UW2 will be represented by "John". In this case, we can hold an **agt** relation between **UW1** and **UW2**.

3.2.3 Attributes: Expressing Subjectivity of the Speaker.

Attributes are mainly used to describe the subjectivity of sentences. They show what is said from the speaker's point of view: how the speaker views what is said. This includes phenomena technically called "speech acts", "propositional attitudes", "truth values", etc. Attributes are used to describe logicality of UWs, *times with respect to the speaker*, speaker's view on aspects of event, speaker's view of reference to concepts, speaker's view of emphasis, focus and topic, speaker's attitudes, speaker's feelings and judgments and attributes for convention.

4. Building the Arabic Dictionary

A UNL dictionary stores information for a language. It stores information concerning what kinds of UWs (concepts) the language expresses and where those words can be used. A word dictionary stores the following items:

1) Universal words for identifying concepts
2) Word headings for universal words that can express concepts
3) Information on the linguistic behavior of words

A word dictionary provides information for computers to understand natural language, and express information in natural language. A dictionary entry consists of a correspondence between a concept and a word, and information concerning morphological and syntactic properties of a word when that correspondence was established.

Each entry in the dictionary has the following format:

[HW] {ID} "UW" (ATTR,…) <FLG,FRE,PRI>;

For example: [ولد] {1} "boy(icl>person)" (CommonNoun,Sing,Masc….etc.) <A,0,0>;

In building this dictionary, we considered that the head word will be stem based because this makes the derivation of plural nouns, for example, easier, without the need of another entry in our dictionary to express the plural noun. In fact, the design of the Arabic dictionary depends entirely on the approach by which the Arabic words have been dealt with. According to our design, the focus of attention is given to the form of the head word of the entry needed to fulfill language analysis and generation tasks adequately. Doing this is twofold: first, it will make it possible to avoid adding all possible inflectional and derivational paradigms of each lexical item to the dictionary (e.g. instead of storing حكومة, حكومتنا , حكومات etc., only حكوم will be stored) (cf. Al-Ansary (2003)). Second, to minimize the number of entries in the dictionary which will give more efficiency in the analysis and generation tasks and minimize the processing time. To reach this target a detailed computational linguistic analysis was conducted on the Arabic word form keeping an eye on both analysis and generation of word forms at the same time, given the fact that the same dictionary should be used in both analysis and generation. Based on this computational linguistic study the best computational form of the lexeme to be stored to represent all its paradigms has been reached.

5. EnConverting Arabic to UNL

The logical structure of Arabic EnConversion process starts by extracting concepts represented by the words of the sentence, and then link these concepts together to form a semantic network of binary relations (the UNL expression). Therefore, the design of the Arabic EnConversion rules is divided into two stages: the morphological analysis stage followed by the 'relation' stage.

In the morphological stage the grammar performs morphological analysis of words in order to 1) extract the correct UWs and 2) assign each UW possible attributes if needed. A number of challenges is faced; this is represented in the segmentation of words into morphemes (words maybe divided different ways), in identifying each morpheme (each morpheme may have more than one function) and in choosing the correct UW (each morpheme may represent more than one concept (UW)). After the choice of UWs has been made, the relation stage starts by assigning semantic relations between them. Different types of information have been stored with UWs that enable the rules to perform correctly in linking correct concepts together and in detecting what the semantic relation needed to link between them is. The following subsections will deal with the morphological and relation stages respectively in more details.

5.1 The Morphological Analysis Stage.

The Arabic language is formally written using characters and diacritics. There are 3 basic diacritics in Arabic, namely, fatha (a), kasra (i) and damma (ou). Which have direct impact on the linguistic interpretation of Arabic words. Although Arabic should be written with full diacritics to avoid misinterpretations, or with mandatory diacritics to minimize ambiguities, most of written Arabic text, except for the religious domain, lack full diacritics. A "human" reading Arabic text is performing contextual analysis in permanence, and sometimes even backtracking in order to reach the correct interpretation of each word and hence the right diacritics to be applied to the word prior to pronouncing it. Therefore, in reading Arabic text, the application of "reading" rules is the easiest part, and unlike Latin languages, must be preceded by analysis, disambiguation and interpretation. Just to feel the task, read the following English sentence: "jst t fl th tsk, rd th fllwng nglsh sntnc", which is the non-vowelized version of the last sentence, in which "fl" could represent (file/foil/fool/feel/fly/flee) and "rd" could represent (rod/road/read/red/raid/ ride) and so on. Therefore, computational processing of the Arabic language is considered as more complex than its Latin counterpart, considering only the issue of absence of diacritics, which increases significantly the amount of morphological and lexical ambiguities resulting in a chain of syntactic ambiguities.

5.1.1 Overview of Arabic Morphology.

The Arabic language has a very rich morphology. Words in Arabic are constructed out of prefixes, stem, infixes and suffixes. The stem itself is composed of two basic elements: the root and the morphological pattern. The root could be a "tri root", consisting of three characters, or a "quad root" composed of four characters. The application of a morphological pattern to a root generates a stem, which is the basic form of an Arabic word token. This demonstrates that Arabic displays both concatenative (linear) and non-concatenative (non-linear) morphology (for non-linear approaches to morphology in general cf. Kiraz (1996) and Koskenniemi (1983), and to Arabic Morphology cf. Kiraz (2001)). The Arabic language consists of about 6000 roots and 700 morphological patterns. Not all patterns could be applied to a given root. The actual valid root-pattern combinations in Arabic generate around 150,000 stems. Two thirds of such stems are considered classical Arabic and only the remaining 50,000 stems are those actually used in daily life.

As a Semitic language, Arabic verb formation is based on either triconsonantal or quadrilateral roots, which is not a word in itself but contains the semantic core. The consonants ك – ت - ب, for example, indicate 'write'. Words are formed by combining the root with a vowel structure (morphological pattern) and with affixes. Arabic verbs can be divided according to denudation and augmentation. (Figure (6)) or according to strength and weakness (Figure (7)).

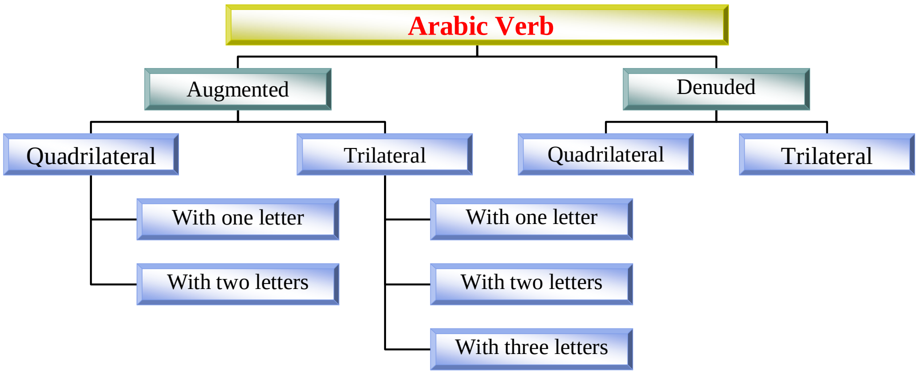


Figure (6): classification of Arabic verbs according to denudation and augmentation.

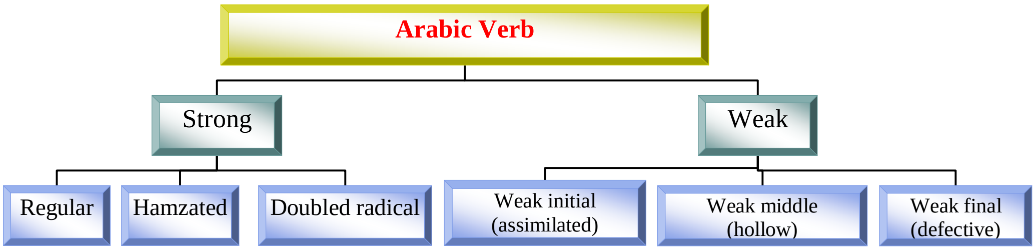


Figure (7): classification of Arabic verbs according to strength and weakness.

Nouns on the other hand, can be divided into solid and derivational. Derivational nouns are those nouns that have verbs from which they are derived. Solid nouns are those nouns that do not have such verbs. Nouns have different morphological attributes that affect their appearance in different sentence structures. Nouns (and their modifying adjectives) are either definite or indefinite (there is an article for the definite state only). A noun is definite if it has the definite article prefix (al-), if it has a suffixed pronoun (kitaabu-ha l-gamil 'her nice book'), if it is inherently definite by being a proper noun, or if it is in a genitive construction (Idafa) with a definite noun or nouns (kitaabu l-benti, 'the book of the girl').

Arabic stems allow attachment of a wide range of prefixes and suffixes. Prefixes in Arabic, for example, could be prepositional (ل,ب, ك), coordinative (ف, و), adverbial (ف), interrogative (أ) and to express future tense (س). Suffixes have a much wider range of functions; for example, they can be used to express subject, object, possession, case (nominative/ accusative/ genitive), number (dual/ proper plural), gender (masculine/ feminine) and person (first/ second/ third).

5.1.2. A Linear Approach to Arabic Morphology.

As Arabic displays a wide range of inflection and derivation, it gives rise to a large space of morphological variation. The UNL formalism is designed to segment any Natural Language input according to the morphemes stored in the dictionary. This means that UNL deals with any input as a sequence of morphemes (linearly)[3]. It can not deal with the derivational side of Arabic in a two-level approach (root + morphological pattern) (cf. Al-Ansary (2004b)). Consequently, it is not possible to derive a word from a root although the nature of Arabic morphology is non-linear. Therefore, both of inflectional and derivational aspects of Arabic should be dealt with concatenatively to be able to adapt Arabic to UNL.

5.1.3. Formalizing Morphological Rules:

In order to make it possible for the morphological component to segment and identify the sequence of morphemes in the input, morphological rules are classified into four groups:

a) The first group

This group focuses on segmenting and identifying prefixes. It composes prefixes to their stems.

Example (1): Composing prefixes attached to verbs

```
- {T1,mor}{21,AUG,^&@past,^TA,mor:&@past,TA}P252;
- {   ا   } {                    ستعد                    }P252;
```

This rule is used to compose verb prefixes like "ا" which has the feature (T1) to the stem "ستعد" which has the features (21, AUG) by giving the UNL attribute @past. When the rule applies, it gives the feature TA to prevent applying the rule more than once on the same node. The feature 'mor' is used to block applying the rule outside the morphological phase.

It could be the case that a given attribute should be added to a given node without the existence of any prefix. For example, the verb "كتب" 'wrote' does not have a prefix to refer to the tense. One possible solution is to rely on the information stored in the dictionary beside other formal cues: either to be preceded by a blank space or to be in the SHEAD node position. See example (2):

[3] For more detailed information on Arabic linear models cf. Beesley et al (2003).

Example (2): Adding Verb attribute (@past) without the existence of any prefix

```
:{^A1,SHEAD,mor}{21,^AUG,^&@past,^TA,mor:&@past,TA}P252;
:{      SHEAD     }{              كتب                    } P252;

:{^A1,blk,mor}{21,^AUG,^&@past,^TA,mor:&@past,TA}(^21)P252;
:{      قد      }{              كتب               }(رواية)P252;
```

Note that although these rules insert the attribute (@past) to verbs like “كتب” ‘wrote’, but it could be a case of an undesired analysis. It could be possible that “كتب” should be interpreted as a plural form of the noun “كتاب”. But at this point the grammar can not decide which concept this word represents. . If any analysis is rejected, backtrack rules will deal with the situation later on in the grammar. The same process occurs with nouns, once the grammar has a possibility of a given noun stem preceded by a nominal prefix, see example (3)

Example (3): Composing prefixes attached to nouns:

```
-{A1,PN,mor}{ST,19,mor:&@def}P252;
-{      ال    }{      معهد          }P252;
```

According to this rule, if the definite article “ال” is attached to any noun stem, it will be composed to the noun giving it the UNL attribute @def.

The rule in example (4) is the same as that in example (3), however it does not give the same attribute (@def) because the stem is an adjective not a noun.

Example (4): Composing prefixes attached to adjectives:

```
-{A1,PN,mor}{ST,22,mor:DEF}P252;
-{      ال      }{      وطني       } P252;
```

Note that adjectives will take the feature DEF to refer to the agreement between the noun and the adjective that follows. It is very important at this point to differentiate between UNL attributes and linguistic features used in the grammar. For example, there is an essential difference between “@def” and “DEF”. The former is an attribute that will be appear in the final UNL expression, the hyper semantic network, but “DEF” is a pure linguistic feature used by means of the author of the grammar to restrict the application of rules.

It may occur that the definite article is preceded by another prefix like a preposition “ل” and “ب”. The rule in example (5) deals with the situation.

Example (5): Dealing with more than one prefix

```
-{PN,A1,LOC,^DIRC,mor}{ST,mor:&@def,PP,loc}P252;
-{          بال                }{        ملعب                }P252;
```

This rule states that if the sequence “بال” opens a word, it could be interpreted as (ال + ب). Accordingly, the node under investigation will take two pieces of information. The first is the UNL attribute “@def” and the second is the linguistic feature “PP” (prepositional phrase) and this prepositional phrase is a locative (loc). Both of PP and ‘loc’ are used later on in building semantic relations between concepts.

b) The second group:

This group of rules deals with suffixes. Left composition rules are used to compose suffixes to their stem. This group of rules applies in two successive phases. The first phase is during the segmentation of the input into morphemes according to the dictionary while the second one is after segmentation has been accomplished.
During the first phase, the segmentation phase, suffixes of nouns and adjectives have been dealt with. Consider the following rule example:

Example (6): Composition during the segmentation phase:

```
+{ST,19,^8,mor}{A2, A2.2,N2,R1.1,mor:&@pl,FEM}P252;
+{     اجتماع     }{                    ات                    }P252;
```

The rule in example (6) states that if a plural feminine suffix is attached to a given nominal stem, the feature “FEM” will be assigned to this stem. Also, the UNL attribute “@pl” (plural) will be assigned to this stem. In the case of an adjectival stems, the node representing this stem will be assigned the feminine feature (FEM) but not the UNL attribute (@pl). See the rule in example (7):

Example (7): Composition suffixes of adjectives during the segmentation phase:

```
+{ST,22,mor}{A2, A2.2,N2,R1.1,mor:FEM}P252;
+{   جميل     }{          ات                    } P252;
```

During the second phase, the after-segmentation phase, some nodes may be considered unnecessary and therefore, it will be marked for deletion. Consider the following rule in example (8):

Example (8): Composition suffixes after segmentation:

```
+(FEM){21,fix}{O,R1.1,fix}P252;
+(قصة){ ألف  }{    ها      } P252;
```

This rule states that if a verb is followed by an object feminine pronoun and this verb is preceded by a feminine noun, this could mean that this object pronoun is referring to this feminine noun, so this pronoun is no longer needed and will be deleted later.

c) The third group:

This group of rules is used to insert nodes that are implicitly expressed in the Arabic input. For example, when a verb like “ضرب” ‘hit’ occurs in a sentence without an agent, the node “هو” will be inserted to play the implicit role of the agent. See the following rule in example (9):.

Example (9): Node insertion rule.

```
:(SHEAD){21,NMAGT,^>agt,fix,^#INSE:#INSE}"[هو]:#INSE,blk,fix"(FS)P255;
:(SHEAD){            ضرب                          }"            هو        "( . )P255;
```

This rule inserts a node for the pronoun [هو] because the left node is marked by a masculine agent given the conditions a) the left node should be preceded by the SHEAD and the right node should be followed by a final stop.

d) The fourth group:

This group of rules deals with deleting unnecessary nodes. This type of rule works in the after-segmentation phase. See the following example:

Example (10): Deleting unused nodes after segmentation.

```
DR{SHEAD}{RE}P255;
DR{SHEAD}{ و }P255;
```

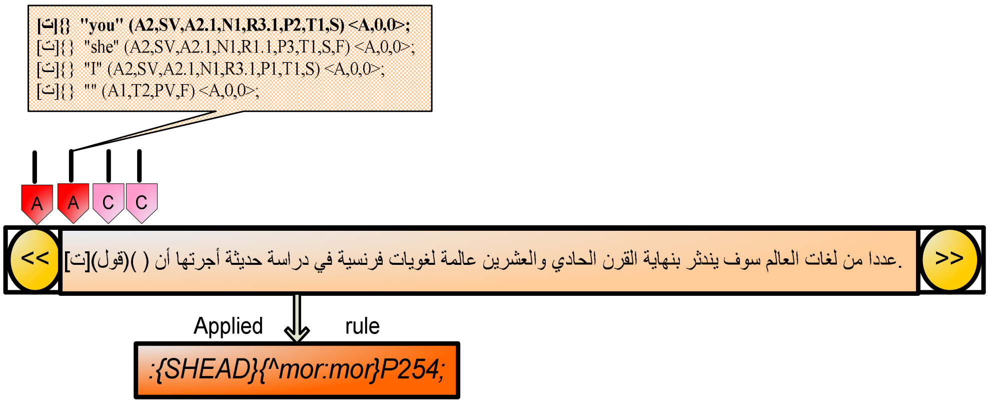
As the UNL system does not make relations between sentences; it deals with sentence by sentence, therefore if the prefix “و” opens a sentence, this means that it can not be used to conjoin two concepts and will be deleted as it does not have any role in assigning any semantic relation.

5.1.3.1 A corpus-based example of morphological analysis:

In this subsection a concrete example will be dealt with to make the morphological analysis clear, the first stage in EnConverting Arabic texts to UNL. This example is one of the sentences taken from our running corpus.
Example:

تقول عالمة لغويات فرنسية في دراسة حديثة أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين.[4]

The following list of figures traces morphological processing, step by step, until the output of this stage has been released. Processing starts by examining the input starting from the initial state and the first morpheme detected by means of the dictionary. See figure (8) below:

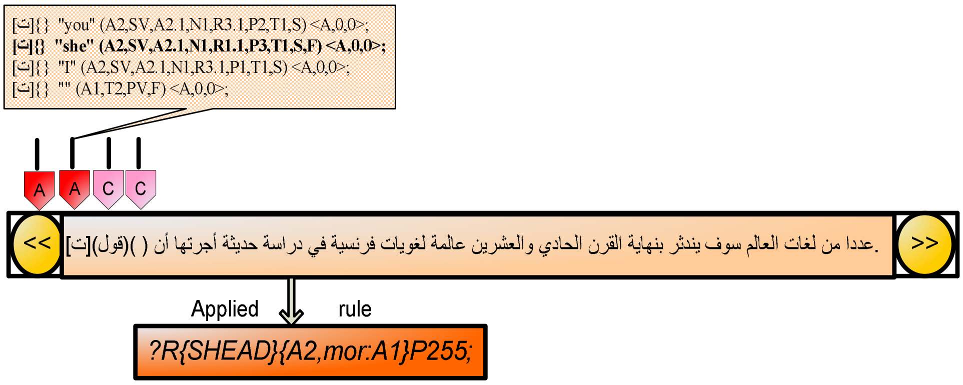


```
[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;
```

A
A
C
C
<<
[ت](قول)( ) أن أجرتها حديثة دراسة في فرنسية لغويات عالمة والعشرين الحادي القرن بنهاية يندثر سوف العالم لغات من عددا.
>>
Applied rule
:{SHEAD}{^mor:mor}P254;

Figure (8)

As shown in figure (8), the left Analysis Window is placed on the Sentence Head (<<) and the extreme right node, namely the first morpheme of the input (ت). The right node is marked for carrying out morphological analysis by the feature 'mor'. This allows to morphological rules to apply on the morpheme (ت). As it is clear from figure (11), the morphological rules examine the first possibility for the morpheme in the right Analysis Window. It tests if it can represent the concept 'you' (this is denoted by the bolded line in the box representing possible targets of the morpheme under investigation). However, one of the morphological rules (backtrack rules) rejected this interpretation as this concept ('you' realized by the morpheme (ت)) can not occur in Arabic as a prefix (Figure (9)).



```
[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;
```

A
A
C
C
<<
[ت](قول)( ) أن أجرتها حديثة دراسة في فرنسية لغويات عالمة والعشرين الحادي القرن بنهاية يندثر سوف العالم لغات من عددا.
>>
Applied rule
?R{SHEAD}{A2,mor:A1}P255;

Figure (9)

The next possibility is under testing, to consider the morpheme (ت) representing the concept 'she'. Again, this possibility fails as this morpheme will be considered as a suffix. The backtrack rule in figure (9) will be repeated until it succeeds with the fourth possibility to consider the morpheme (ت) as a prefix (Figure (10)).

[4] This sentence is taken from Al-Ahram newspaper, Thursday 19 January 2006, Issue number 43508.

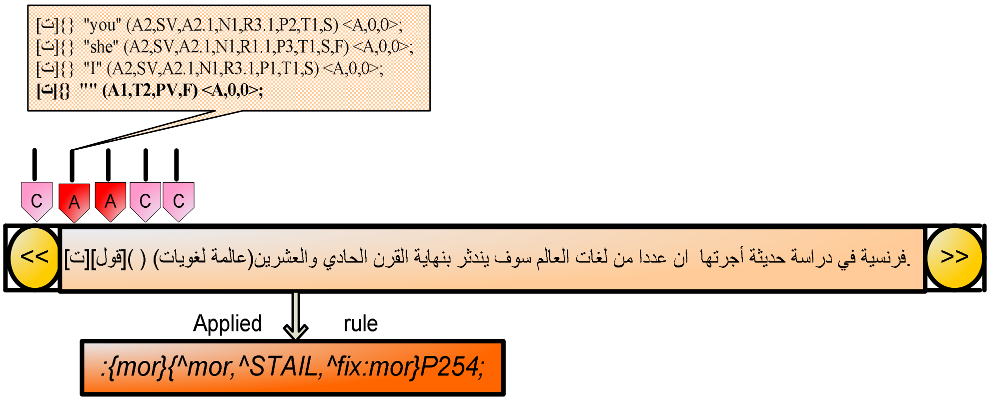

[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;
C
A
A
C
C
<<
فرنسية في دراسة حديثة أجرتها ان عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين(عالمة لغويات) ( )[قول][ت].
>>
Applied
rule
:{mor}{^mor,^STAIL,^fix:mor}P254;

Figure (10)

In figure (10), the next possible morpheme detected by means of the dictionary is (قول). It will be given the feature 'mor' to permit morphological rules to apply on this node. In figure (11), the dictionary suggests that the right hand window could be interpreted as a form of "قال" in the present tense.

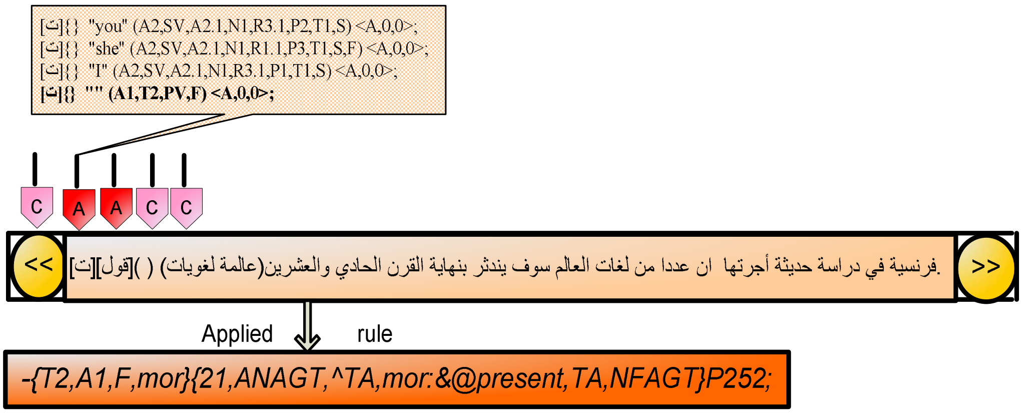

[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;
C
A
A
C
C
<<
فرنسية في دراسة حديثة أجرتها ان عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين(عالمة لغويات) ( )[قول][ت].
>>
Applied
rule
-{T2,A1,F,mor}{21,ANAGT,^TA,mor:&@present,TA,NFAGT}P252;

Figure (11)

In figure (11), once the EnConverter found suitable selections for the left and right analysis windows, a right composition rule will be apply to compose the verb prefix to the verb stem. Having reached this, the right analysis window will be assigned the UNL attribute @present and will be marked for a feminine agent.

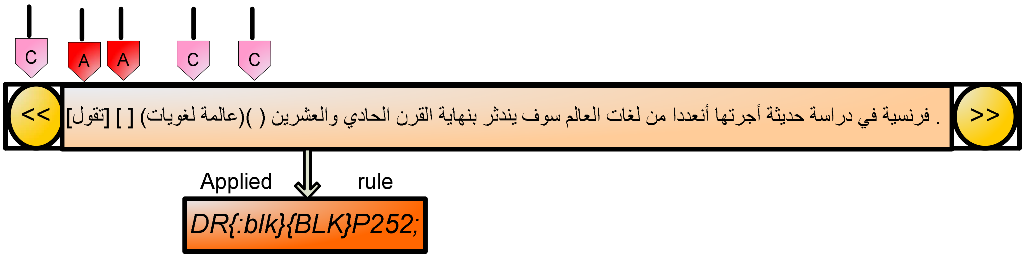

C
A
A
C
C
<<
فرنسية في دراسة حديثة أجرتها أنعددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين ( )(عالمة لغويات) [ ] [تقول] .
>>
Applied
rule
DR{:blk}{BLK}P252;

Figure (12)

In figure (12), the right analysis window moves rightward and detects a blank follows. The EnConverter detects blanks automatically and assigns them the feature "BLK" without the need to define them in the dictionary. Blanks are word boundaries , therefore, the right node deletion rule (the applied rule in figure (12)) is carried out to delete blank spaces giving the feature "blk" to the word that precedes.

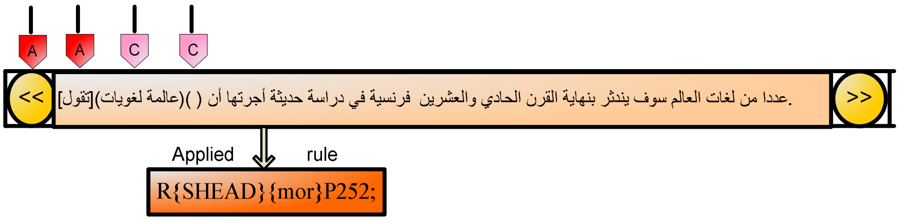

A
A
C
C
<<
عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين فرنسية في دراسة حديثة أجرتها أن ( )(عالمة لغويات)[تقول].
>>
Applied rule
R{SHEAD}{mor}P252;

Figure (13)

After the deletion rule has been carried out, the EnConverter moves backwards to check if there is any rule that can be applied on the new composed nodes. As there is no rule can be applied on the new nodes, as seen in figure (13), a right shift rule is applied to move the analysis windows to the right.

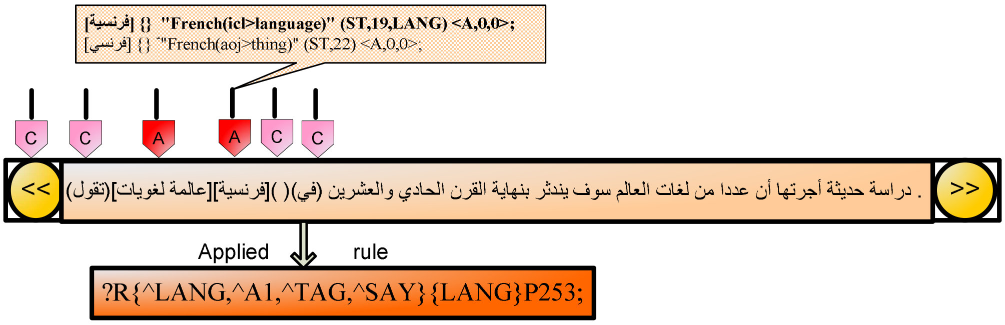

[فرنسية] {} "French(icl>language)" (ST,19,LANG) <A,0,0>;
[فرنسي] {} "French(aoj>thing)" (ST,22) <A,0,0>;
C
C
A
A
C
C
<<
. دراسة حديثة أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (في)( )[فرنسية][عالمة لغويات](تقول)
>>
Applied rule
?R{^LANG,^A1,^TAG,^SAY}{LANG}P253;

Figure (14)

As seen in figure (14), the dictionary retrieves two possible analyses for the word “فرنسية”. The first is “فرنسية” as one morpheme representing the concept “French language”, while the second is “فرنسي” + “ة” as two morphemes representing an adjectival concept + a feminine suffix respectively. The grammar tries the first possibility to consider “فرنسية” a morpheme representing the concept “French language”. However, this possibility fails as, according to our consulted corpus, this concept is most probably preceded by a say verb or by the morpheme “لغة”. Therefore, the backtrack rule in figure (14) tries to find another solution. In figure (15), the other alternative is tried: to consider the right Analysis Window which contains a morpheme that represents an adjectival concept followed by a feminine suffix. As the dictionary provides that the left analysis window represents a feminine concept, the second analysis is preferred.

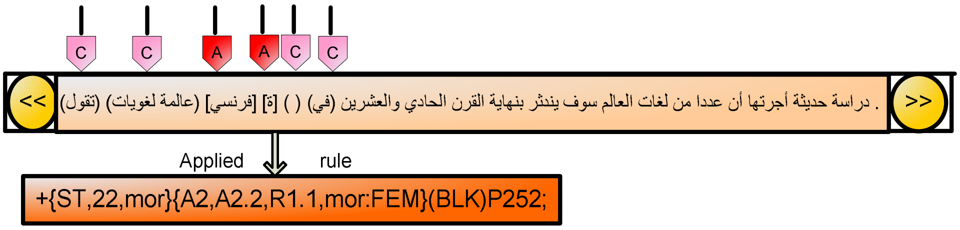

C
C
A
A
C
C
<<
. دراسة حديثة أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (في) ( ) [ة] [فرنسي] (عالمة لغويات) (تقول)
>>
Applied rule
+{ST,22,mor}{A2,A2.2,R1.1,mor:FEM}(BLK)P252;

Figure (15)

Accordingly, the applied left composition rule in figure (15) is carried out to compose the feminine suffix “ة” to the adjective “فرنسي”. In figure (16), the EnConverter moves right and stops by a new node which is “في”. The engine retrieves possible interpretations for the new node.

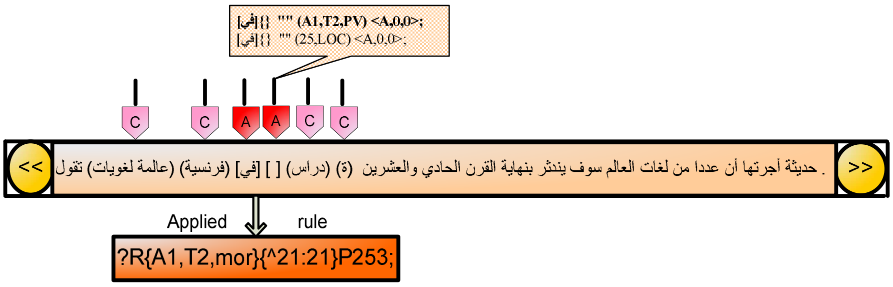

[في]{} "" (A1,T2,PV) <A,0,0>;
[في]{} "" (25,LOC) <A,0,0>;
C
C
A
A
C
C
<<
. حديثة أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (ة) (دراس) [ ] [في] (فرنسية) (عالمة لغويات) تقول
>>
Applied rule
?R{A1,T2,mor}{^21:21}P253;

Figure (16)

The EnConverter tries the first alternative for “في” which is a prefix expressing the present tense and needs a verb (in the present tense) after it, so the back track rule for the right node in figure (16) rejects this situation because the node following this verb prefix is a blank space not a verb.

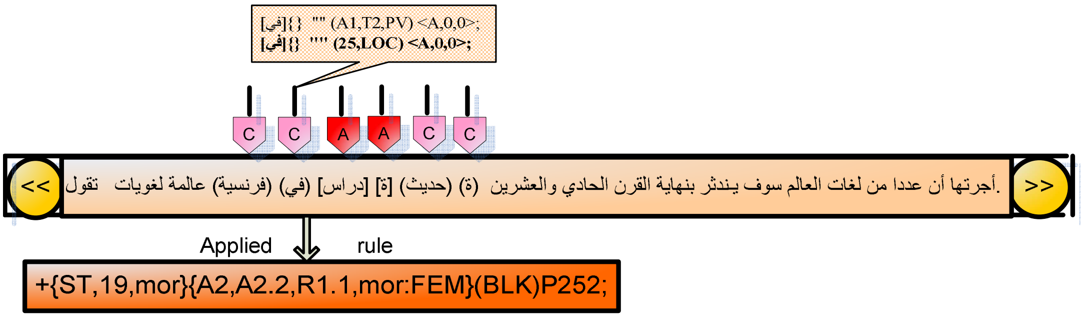

[في]{} "" (A1,T2,PV) <A,0,0>;
[في]{} "" (25,LOC) <A,0,0>;
C
C
A
A
C
C
<<
أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (ة) (حديث) [ة] [دراس] (في) (فرنسية) عالمة لغويات تقول
>>
Applied
rule
+{ST,19,mor}{A2,A2.2,R1.1,mor:FEM}(BLK)P252;

Figure (17)

After selecting the suitable interpretation for the morpheme “في”, the EnConverter moves to the next nodes, namely the “دراس” and “ة” (figure 17). The two nodes are interpreted as a noun followed by a feminine suffix, therefore, the suffix is composed to its stem by the applied left composition rule. In this case the analysis windows will be moved backwards to check if there is any rule which can be applied on the two nodes after composition (figure (18)).

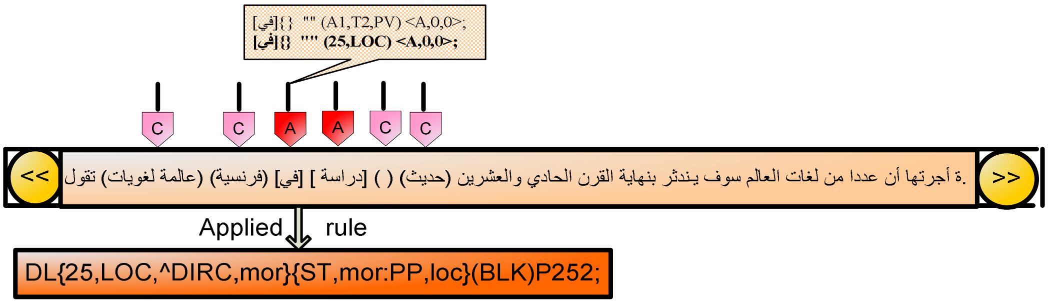

[في]{} "" (A1,T2,PV) <A,0,0>;
[في]{} "" (25,LOC) <A,0,0>;
C
C
A
A
C
C
<<
ة. أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (حديث) ( ) [دراسة] [في] (فرنسية) (عالمة لغويات) تقول
>>
Applied
rule
DL{25,LOC,^DIRC,mor}{ST,mor:PP,loc}(BLK)P252;

Figure (18)

In figure (18), the EnConverter found out that there is a rule that can be applied on the previous nodes. A deletion rule is applied to delete the left node leaving a trace on the right node marking it as a start of a Prepositional Phrase (PP) and that the deleted preposition is locative (loc).

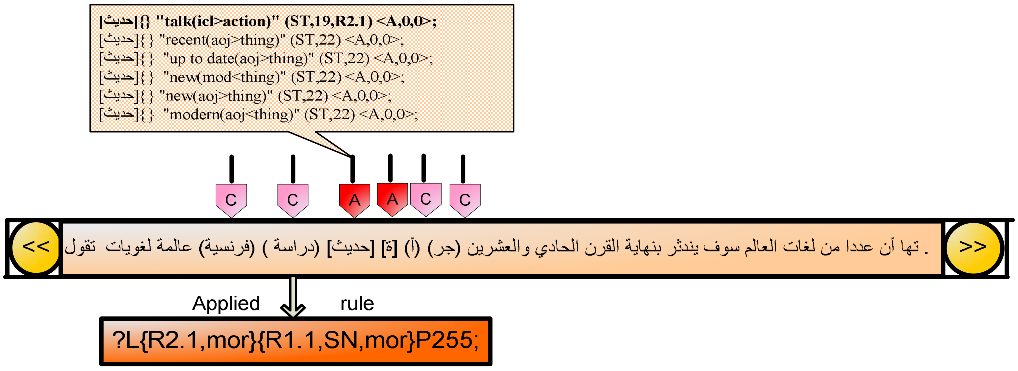
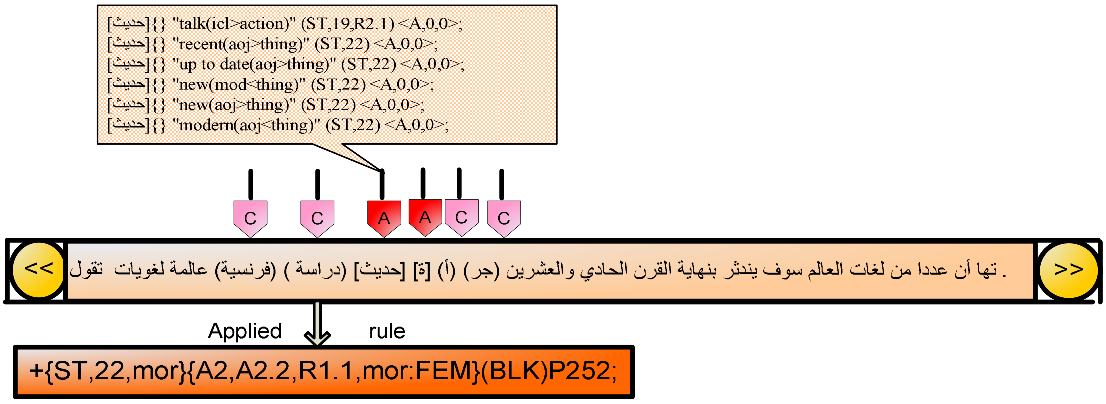

[حديث]{} "talk(icl>action)" (ST,19,R2.1) <A,0,0>;
[حديث]{} "recent(aoj>thing)" (ST,22) <A,0,0>;
[حديث]{} "up to date(aoj>thing)" (ST,22) <A,0,0>;
[حديث]{} "new(mod<thing)" (ST,22) <A,0,0>;
[حديث]{} "new(aoj>thing)" (ST,22) <A,0,0>;
[حديث]{} "modern(aoj<thing)" (ST,22) <A,0,0>;
C
C
A
A
C
C
<<
. تها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (جر) (أ) [ة] [حديث] (دراسة) (فرنسية) عالمة لغويات تقول
>>
Applied
rule
?L{R2.1,mor}{R1.1,SN,mor}P255;

Figure (19)

Having finished with “دراسة”, the EnConverter moves right and retrieves from the dictionary the possible entries for the next possible node “حديث” (figure 19). “talk” is not a possible interpretation of the left node because it is masculine noun that can not be feminine by adding the feminine suffix “ة”. Therefore the backtrack rule in figure (19) applies to find another solution for “حديث” that supports a following feminine suffix.



```
[حديث]{} "talk(icl>action)" (ST,19,R2.1) <A,0,0>;
[حديث]{} "recent(aoj>thing)" (ST,22) <A,0,0>;
[حديث]{} "up to date(aoj>thing)" (ST,22) <A,0,0>;
[حديث]{} "new(mod<thing)" (ST,22) <A,0,0>;
[حديث]{} "new(aoj>thing)" (ST,22) <A,0,0>;
[حديث]{} "modern(aoj<thing)" (ST,22) <A,0,0>;
```

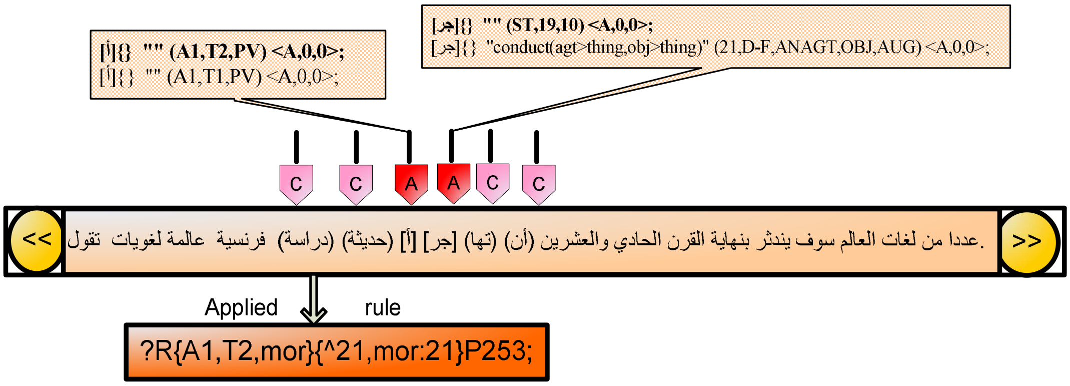
تها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (جر) (أ) [ة] [حديث] (دراسة) (فرنسية) عالمة لغويات تقول .

Applied rule

```
+{ST,22,mor}{A2,A2.2,R1.1,mor:FEM}(BLK)P252;
```

Figure (20)

A suitable solution is the second alternative as shown in figure (20); it is an adjectival concept. This alternative is suitable as the right condition window has a feminine noun. Therefore, it could be possible for this noun to be followed by a feminine adjective representing an adjectival concept. In this case, the rule in figure (20) is applied to compose the suffix "ة" to its stem to form the node "حديثة".



```
[أ]{} "" (A1,T2,PV) <A,0,0>;
[أ]{} "" (A1,T1,PV) <A,0,0>;
```

```
[جر]{} "" (ST,19,10) <A,0,0>;
[جر]{} "conduct(agt>thing,obj>thing)" (21,D-F,ANAGT,OBJ,AUG) <A,0,0>;
```

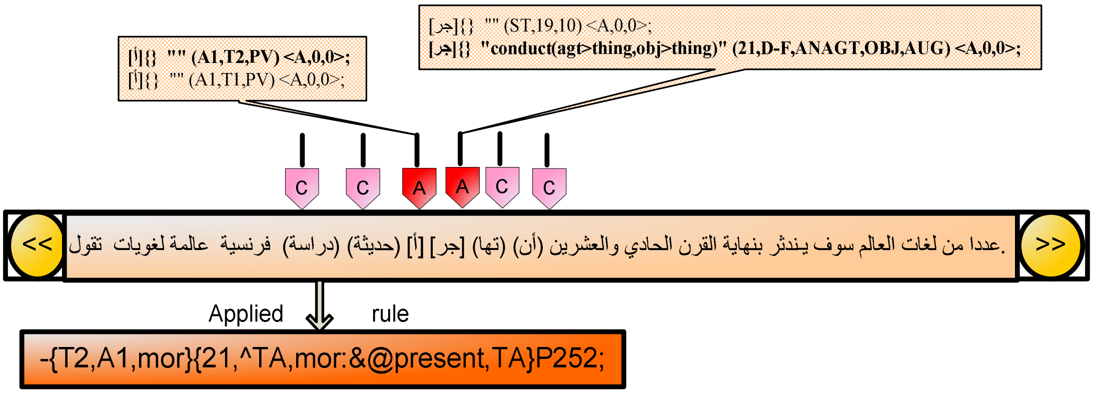
عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (أن) (تها) [جر] [أ] (حديثة) (دراسة) فرنسية عالمة لغويات تقول.

Applied rule

```
?R{A1,T2,mor}{^21,mor:21}P253;
```

Figure (21)

The two analysis windows move to the right and stop by the node "أ" which has two possibilities. The first is a prefix representing the present tense and the second is a prefix representing the past tense. The EnConverter does not have at this point enough information to decide which type of prefix "أ" is. Therefore, the first choice will be selected considering the "أ" as a present tense prefix. As for the right analysis window, the available interpretation is a noun, which does not go with the verbal prefix in the left analysis window. Consequently, the backtrack rule in figure (21) tries to find another alternative for it.



```
[أ]{} "" (A1,T2,PV) <A,0,0>;
[أ]{} "" (A1,T1,PV) <A,0,0>;
```

```
[جر]{} "" (ST,19,10) <A,0,0>;
[جر]{} "conduct(agt>thing,obj>thing)" (21,D-F,ANAGT,OBJ,AUG) <A,0,0>;
```

عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (أن) (تها) [جر] [أ] (حديثة) (دراسة) فرنسية عالمة لغويات تقول.

Applied rule

```
-{T2,A1,mor}{21,^TA,mor:&@present,TA}P252;
```

Figure (22)

According to the above choices in figure (22), the two nodes will be composed and the verb in the right node takes the UNL attribute @present. The Analysis Windows move to the right to examine the next node as shown in figure (23).

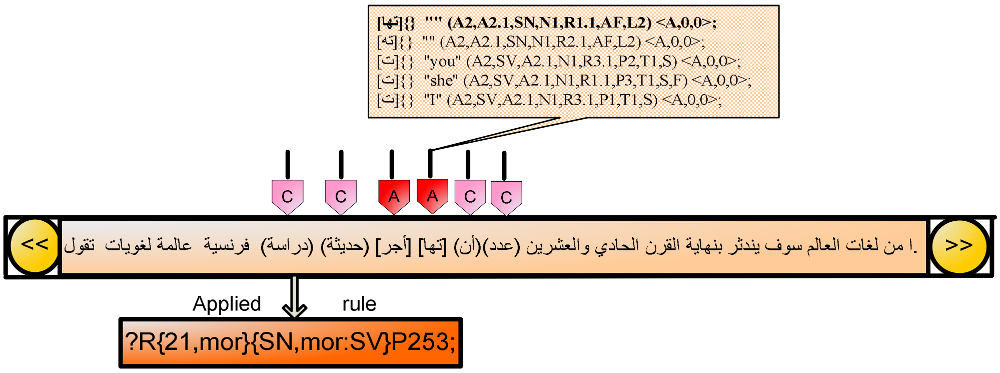

[تها]{} "" (A2,A2.1,SN,N1,R1.1,AF,L2) <A,0,0>;
[ته]{} "" (A2,A2.1,SN,N1,R2.1,AF,L2) <A,0,0>;
[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
C
C
A
A
C
C
<<
.ا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (عدد)(أن) [تها] [أجر] (حديثة) (دراسة) فرنسية عالمة لغويات تقول
>>
Applied rule
?R{21,mor}{SN,mor:SV}P253;

Figure (23)

The first possibility faced is to interpret the right hand node as a nominal suffix. This will be rejected by the backtrack rule in figure (23) as the left node is a verb.

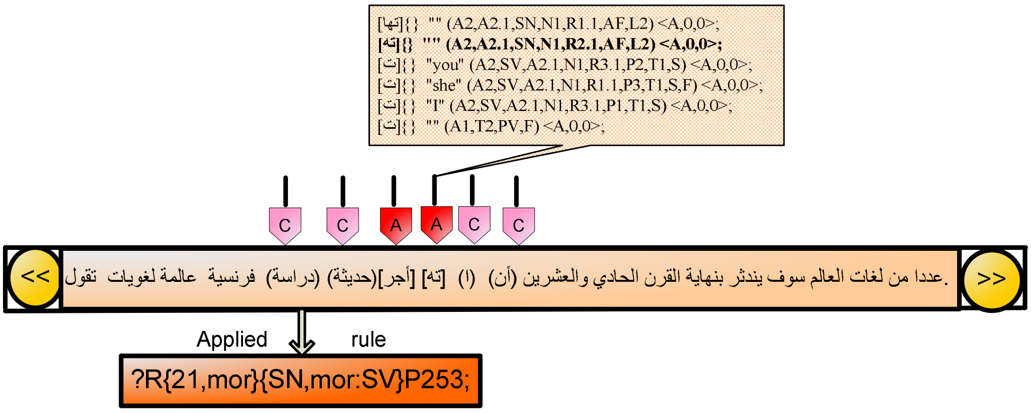

[تها]{} "" (A2,A2.1,SN,N1,R1.1,AF,L2) <A,0,0>;
[ته]{} "" (A2,A2.1,SN,N1,R2.1,AF,L2) <A,0,0>;
[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;
C
C
A
A
C
C
<<
.عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (أن) (ا) [ته] [أجر](حديثة) (دراسة) فرنسية عالمة لغويات تقول
>>
Applied rule
?R{21,mor}{SN,mor:SV}P253;

Figure (24)

Unfortunately, the second alternative for the right analysis window is also another type of nominal suffix which will be rejected by the same backtrack rule applied in figure (23). The EnConverter will try to find another alternative.

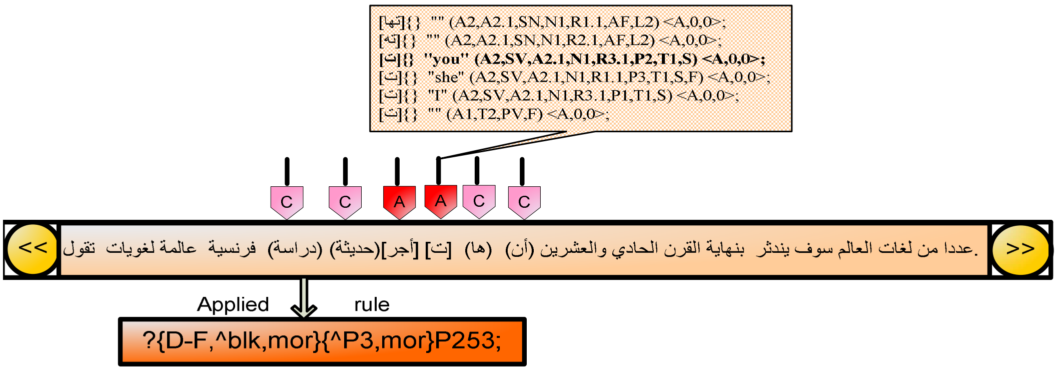
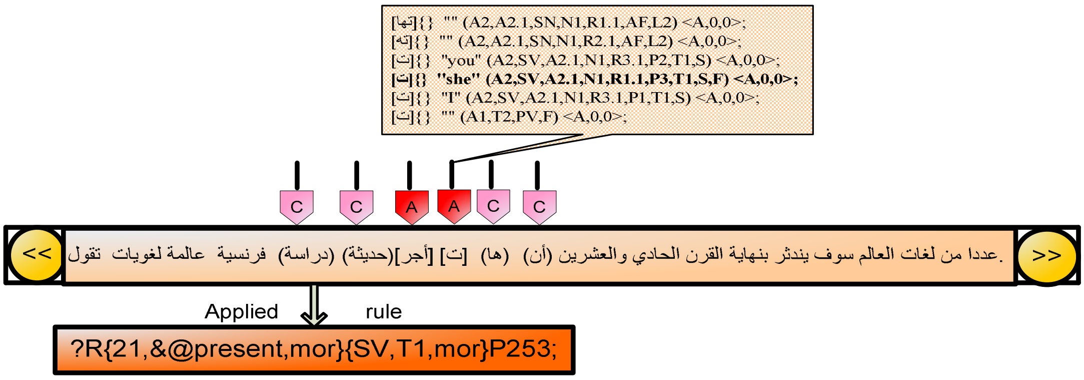

[تها]{} "" (A2,A2.1,SN,N1,R1.1,AF,L2) <A,0,0>;
[ته]{} "" (A2,A2.1,SN,N1,R2.1,AF,L2) <A,0,0>;
[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;
C
C
A
A
C
C
<<
.عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (أن) (ها) [ت] [أجر](حديثة) (دراسة) فرنسية عالمة لغويات تقول
>>
Applied rule
?{D-F,^blk,mor}{^P3,mor}P253;

Figure (25)

Although, the third alternative in figure (25) is a verbal suffix which is rejected because of the attachment of the suffix “ت” ‘you’ to the defective verb should not delete the final weak letter. This suffix should be attached to the form “أجري” not “أجر”, so the back track rule in figure (25) is applied to find another alternative.



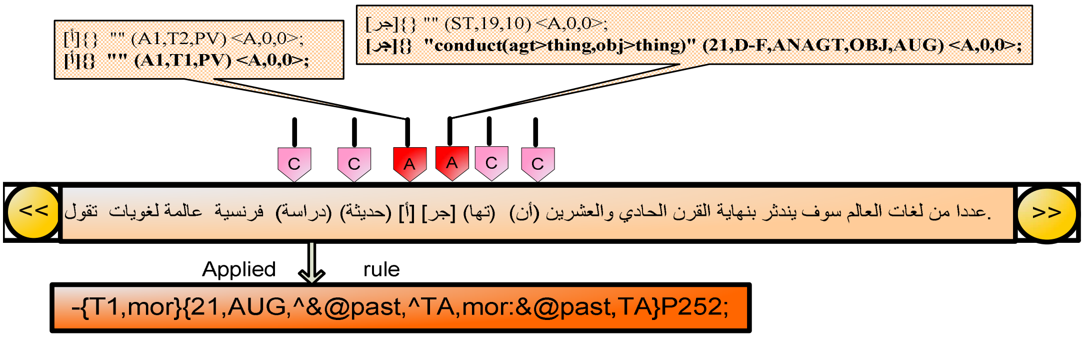
[تها]{} "" (A2,A2.1,SN,N1,R1.1,AF,L2) <A,0,0>;
[ته]{} "" (A2,A2.1,SN,N1,R2.1,AF,L2) <A,0,0>;
[ت]{} "you" (A2,SV,A2.1,N1,R3.1,P2,T1,S) <A,0,0>;
[ت]{} "she" (A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;
[ت]{} "I" (A2,SV,A2.1,N1,R3.1,P1,T1,S) <A,0,0>;
[ت]{} "" (A1,T2,PV,F) <A,0,0>;

عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (أن) (ها) [ت] [أجر](حديثة) (دراسة) فرنسية عالمة لغويات تقول.

?R{21,&@present,mor}{SV,T1,mor}P253;

Figure (26)

After the right node is selected the back track rule in figure (26) will be applied to exchange the verb in the left node which is supposed to be in the present tense because the suffix “ت” ‘she’ is always attached to verbs in the past tense. The back track rule does not find any other verb in the same form in the past tense. Therefore, the “أجر” will be decomposed back into [أ], [جر]. The back track rule will return back to the step in figure (21) and select another alternative for the prefix “أ”, so it will be changed to a prefix representing the past tense not the present as was erroneously decided in figure (21). The new situation is shown in figure (27).



[أ]{} "" (A1,T2,PV) <A,0,0>;
[أ]{} "" (A1,T1,PV) <A,0,0>;

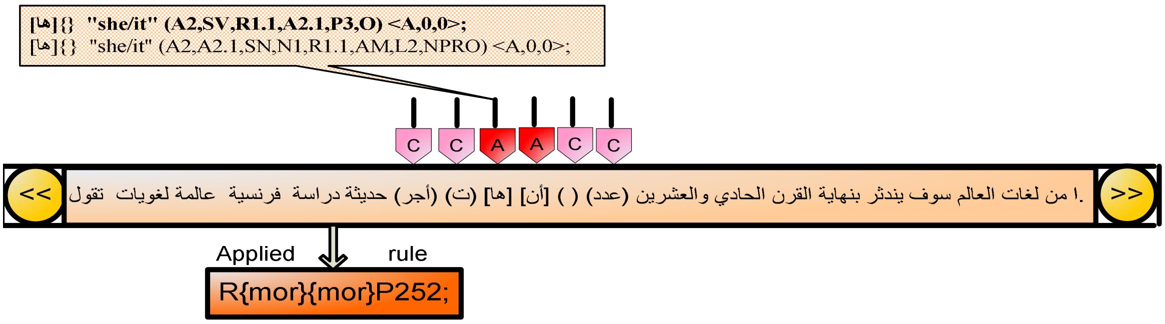
[جر]{} "" (ST,19,10) <A,0,0>;
[جر]{} "conduct(agt>thing,obj>thing)" (21,D-F,ANAGT,OBJ,AUG) <A,0,0>;

عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (أن) (تها) [جر] [أ] (حديثة) (دراسة) فرنسية عالمة لغويات تقول.

-{T1,mor}{21,AUG,^&@past,^TA,mor:&@past,TA}P252;

Figure (27)

After the left node has been exchanged, the right composition rule will be applied and gives the new composed node “أجر” the UNL attribute @past (instead of @present). The EnConverter continues moving to the right till the two analysis windows are standing on [أجر] and [تها]. Steps described in figure (23) will be repeated until the morpheme “ت” is interpreted again as “she” after which the object pronoun “ها” is recognized as “she/it”.



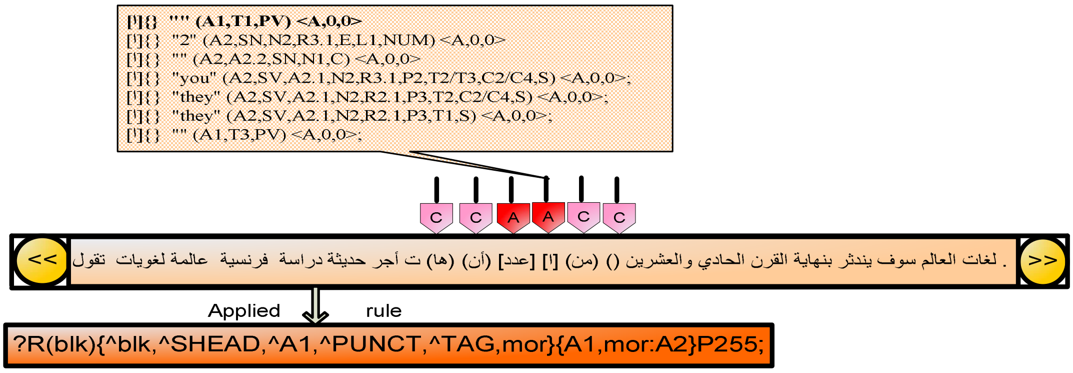
[ها]{} "she/it" (A2,SV,R1.1,A2.1,P3,O) <A,0,0>;
[ها]{} "she/it" (A2,A2.1,SN,N1,R1.1,AM,L2,NPRO) <A,0,0>;

ا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين (عدد) ( ) [أن] [ها] (ت) (أجر) حديثة دراسة فرنسية عالمة لغويات تقول.

R{mor}{mor}P252;

Figure (28)

After the morpheme "ها" has been recognized, the EnConverter didn't find any other morphological rule to apply so, it moves to the right by the right shift rule in figure (28). No rule is detected to apply in this situation which leads the EnConverter to move again applying the same right shift rule till it reaches the node "عدد" in the left analysis window and the node "ا" in the right analysis window (figure (29))



```
[ا]{}  "" (A1,T1,PV) <A,0,0>
[ا]{}  "2" (A2,SN,N2,R3.1,E,L1,NUM) <A,0,0>
[ا]{}  "" (A2,A2.2,SN,N1,C) <A,0,0>
[ا]{}  "you" (A2,SV,A2.1,N2,R3.1,P2,T2/T3,C2/C4,S) <A,0,0>;
[ا]{}  "they" (A2,SV,A2.1,N2,R2.1,P3,T2,C2/C4,S) <A,0,0>;
[ا]{}  "they" (A2,SV,A2.1,N2,R2.1,P3,T1,S) <A,0,0>;
[ا]{}  "" (A1,T3,PV) <A,0,0>;
```

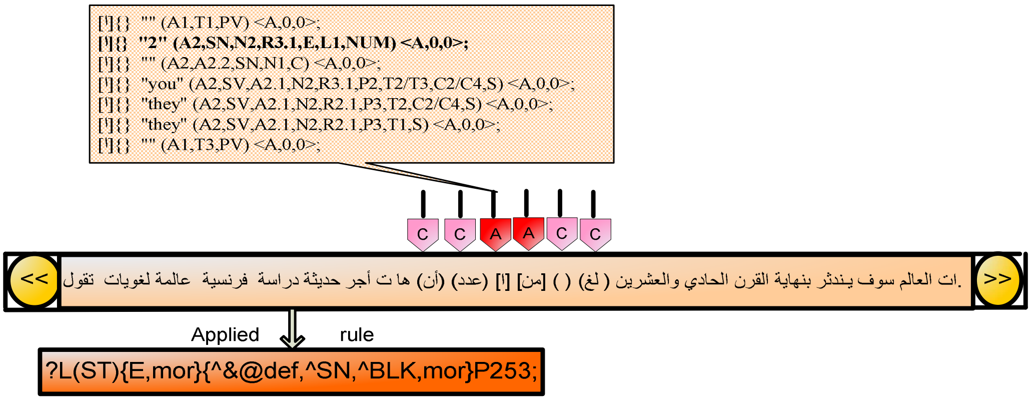
. لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين () (من) [ا] [عدد] (أن) (ها) ت أجر حديثة دراسة فرنسية عالمة لغويات تقول

Applied rule

```
?R(blk){^blk,^SHEAD,^A1,^PUNCT,^TAG,mor}{A1,mor:A2}P255;
```

Figure (29)

Unfortunately, the interpretation of the right node is not suitable for the left node, as it considers the right window as a verbal prefix while the left node is a noun. Consequently, the backtrack rule rejected the right node.



```
[ا]{}  "" (A1,T1,PV) <A,0,0>;
[ا]{}  "2" (A2,SN,N2,R3.1,E,L1,NUM) <A,0,0>;
[ا]{}  "" (A2,A2.2,SN,N1,C) <A,0,0>;
[ا]{}  "you" (A2,SV,A2.1,N2,R3.1,P2,T2/T3,C2/C4,S) <A,0,0>;
[ا]{}  "they" (A2,SV,A2.1,N2,R2.1,P3,T2,C2/C4,S) <A,0,0>;
[ا]{}  "they" (A2,SV,A2.1,N2,R2.1,P3,T1,S) <A,0,0>;
[ا]{}  "" (A1,T3,PV) <A,0,0>;
```

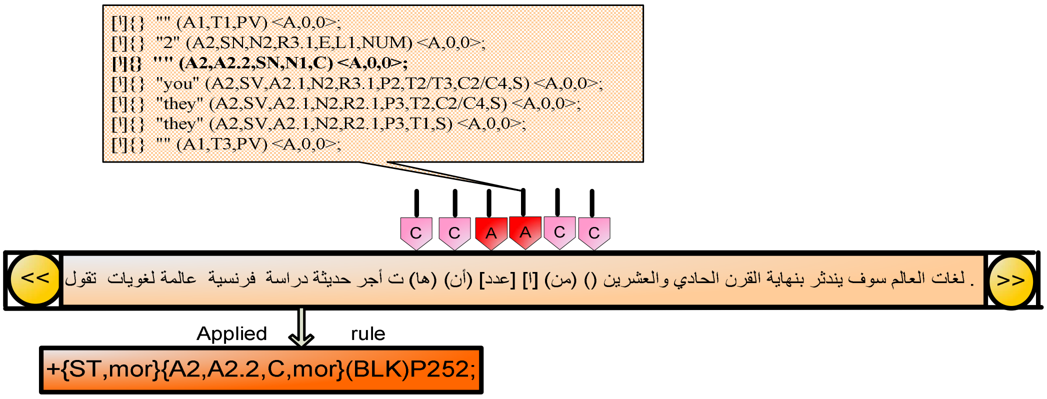
.ات العالم سوف يندثر بنهاية القرن الحادي والعشرين (لغ) ( ) [من] [ا] (عدد) (أن) ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول

Applied rule

```
?L(ST){E,mor}{^&@def,^SN,^BLK,mor}P253;
```

Figure (30)

The other solution that is shown in figure (30) represents a suffix which can be considered temporarily suitable as it is a nominal suffix preceded by a noun. Accordingly, it moves right stopping by “من” in the right node. At this point, the EnConverter, realized that the left node was not a good interpretation as it is followed by a preposition while a noun is expected in this slot after this type of suffixes. The morpheme “ا” which was interpreted as a dual suffix occurs only in case of construct state ‘Idafa’, e.g. the suffix as in "كتابا الطالب" ‘the two books of the student’. According to the new situation in figure (30), the backtrack rule stats to find another interpretation for the previously defined suffix.



```
[ا]{}  "" (A1,T1,PV) <A,0,0>;
[ا]{}  "2" (A2,SN,N2,R3.1,E,L1,NUM) <A,0,0>;
[ا]{}  "" (A2,A2.2,SN,N1,C) <A,0,0>;
[ا]{}  "you" (A2,SV,A2.1,N2,R3.1,P2,T2/T3,C2/C4,S) <A,0,0>;
[ا]{}  "they" (A2,SV,A2.1,N2,R2.1,P3,T2,C2/C4,S) <A,0,0>;
[ا]{}  "they" (A2,SV,A2.1,N2,R2.1,P3,T1,S) <A,0,0>;
[ا]{}  "" (A1,T3,PV) <A,0,0>;
```

. لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين () (من) [ا] [عدد] (أن) (ها) ت أجر حديثة دراسة فرنسية عالمة لغويات تقول

Applied rule

```
+{ST,mor}{A2,A2.2,C,mor}(BLK)P252;
```

Figure (31)

As the EnConverter could find another suitable solution for the suffix "ا", the nunnation suffix, it returned back one step to test if this new interpretation of the right node could have a morphological relation with the previous node ‘عدد’. Fortunately, a composition rule has applied to form a new node "عددا".

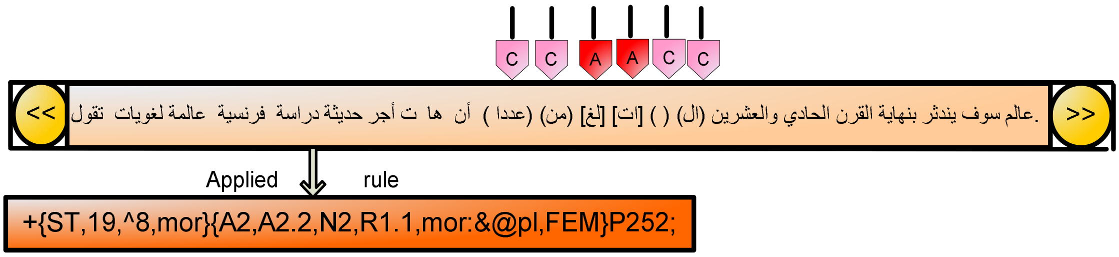

C
C
A
A
C
C
<<
. عالم سوف يندثر بنهاية القرن الحادي والعشرين (ال) ( ) [ات] [لغ] (من) (عددا) أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied
rule
+{ST,19,^8,mor}{A2,A2.2,N2,R1.1,mor:&@pl,FEM}P252;

Figure (32)

In figure (32), the EnConverter moves right to analyze the next morpheme. The possible analysis found for the left node is a noun followed by a plural feminine noun suffix in the right node. The suffix is composed to the nominal stem by means of the composition rule above attaching the UNL attribute “@pl” and the linguistic feature to the noun.

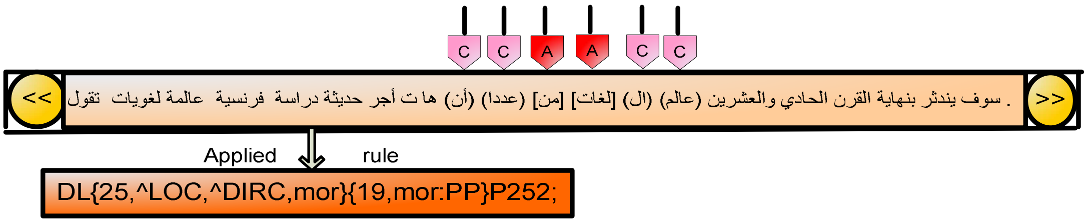

C
C
A
A
C
C
<<
. سوف يندثر بنهاية القرن الحادي والعشرين (عالم) (ال) [لغات] [من] (عددا) (أن) ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied
rule
DL{25,^LOC,^DIRC,mor}{19,mor:PP}P252;

Figure (33)

In figure (33), the EnConverter continues after the last node detected stopping at the preposition. A deletion rule deletes the preposition marking the next node as a start of a prepositional phrase (PP).

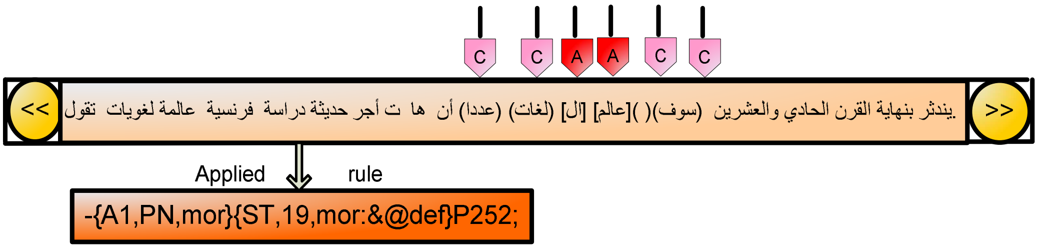
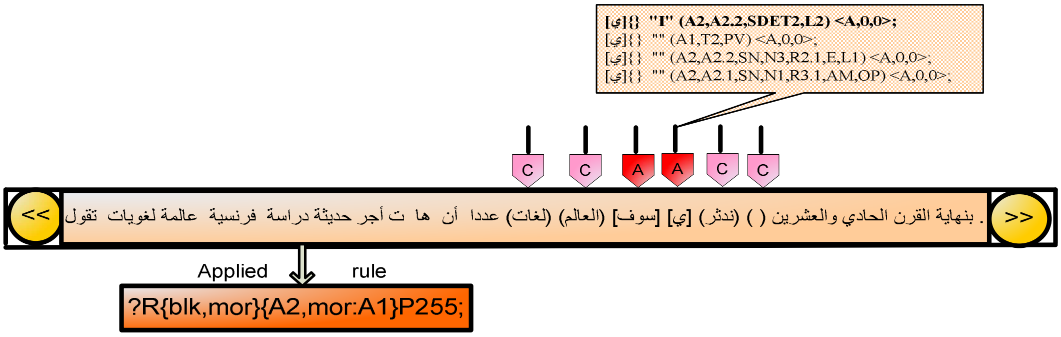

C
C
A
A
C
C
<<
. يندثر بنهاية القرن الحادي والعشرين (سوف)( )[عالم] [ال] (لغات) (عددا) أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied
rule
-{A1,PN,mor}{ST,19,mor:&@def}P252;

Figure (34)

In figure (34) the EnConverter detects a definite article in the left node followed by a noun in the right node. A right composition rule applies to compose this prefix to its stem giving the stem the UNL attribute “@def”.



```
[ي]{} "I" (A2,A2.2,SDET2,L2) <A,0,0>;
[ي]{} "" (A1,T2,PV) <A,0,0>;
[ي]{} "" (A2,A2.2,SN,N3,R2.1,E,L1) <A,0,0>;
[ي]{} "" (A2,A2.1,SN,N1,R3.1,AM,OP) <A,0,0>;
```

C
C
A
A
C
C
<<
. بنهاية القرن الحادي والعشرين ( ) (ندثر) [ي] [سوف] (العالم) (لغات) عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied
rule
?R{blk,mor}{A2,mor:A1}P255;

Figure (34)

As we have illustrated earlier in figure (12), when the blank space is deleted the preceding word takes the feature ‘blk’, accordingly, we can determine word boundaries. Consequently, the right node in figure (35) is rejected to be a suffix as it is preceded by a ‘blk’. The back track rule is dealing with the situation to consider the right node as a prefix instead.

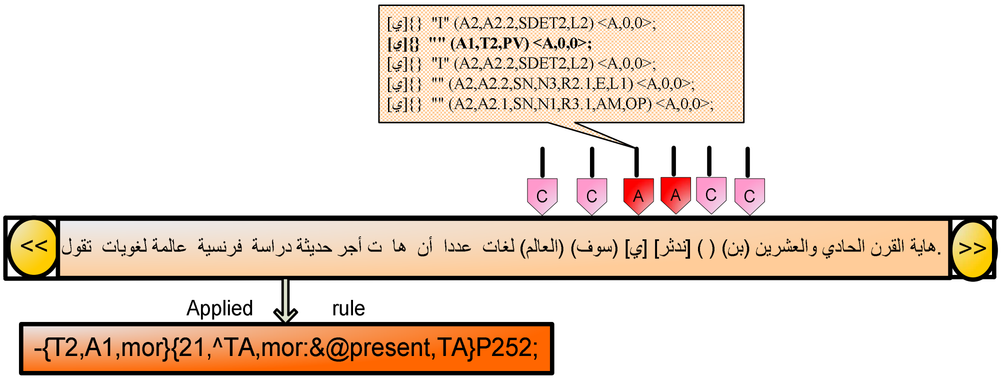

[ي]{} "I" (A2,A2.2,SDET2,L2) <A,0,0>;
[ي]{} "" (A1,T2,PV) <A,0,0>;
[ي]{} "I" (A2,A2.2,SDET2,L2) <A,0,0>;
[ي]{} "" (A2,A2.2,SN,N3,R2.1,E,L1) <A,0,0>;
[ي]{} "" (A2,A2.1,SN,N1,R3.1,AM,OP) <A,0,0>;
C C A A C C
<<
>>
Applied rule
-{T2,A1,mor}{21,^TA,mor:&@present,TA}P252;

Figure (36)

Deciding the left node to be a prefix for the present tense made the interpretation of the right node in figure (36) predictable as a verb. Therefore, the right node composition rule is applied to assign the verb the UNL attribute "@present".

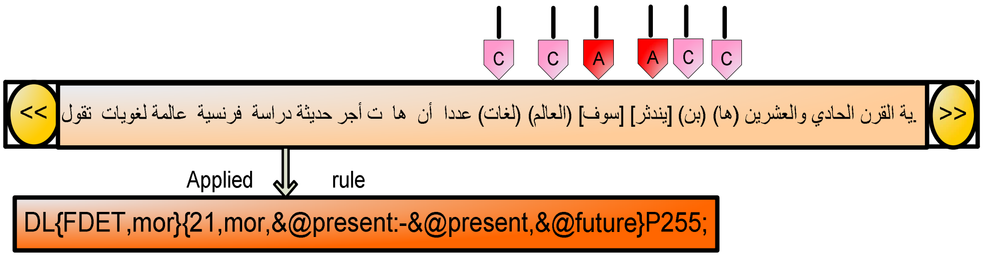


Figure (37)

After the application of composition rules, the EnConverter moves backward to check if there is any other rule that can apply on the composed node. According to figure (37), the EnConverter discovered that a deletion rule can be applied to delete the node "سوف" replacing the UNL attribute "@present" of the right node with "@future".

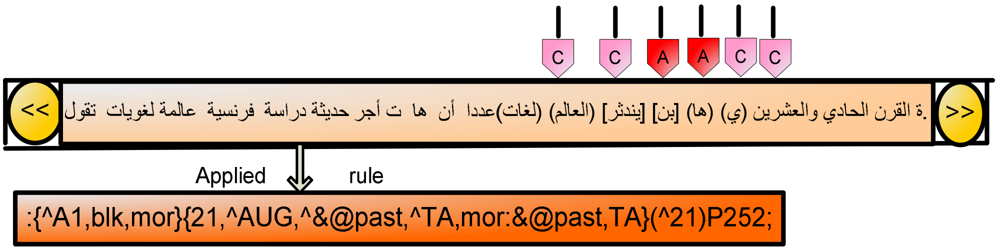


Figure (38)

As the EnConverter is moving, the next possible morpheme is interpreted as a paradigm of the verb "بنى" 'build' giving the right node the UNL attribute "@past" (figure (38)).

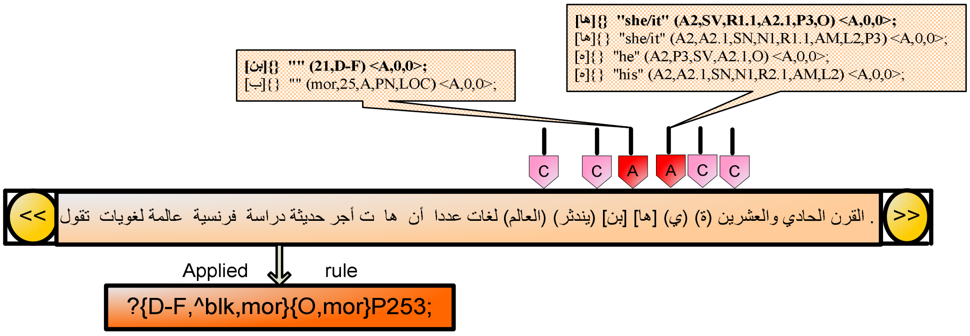
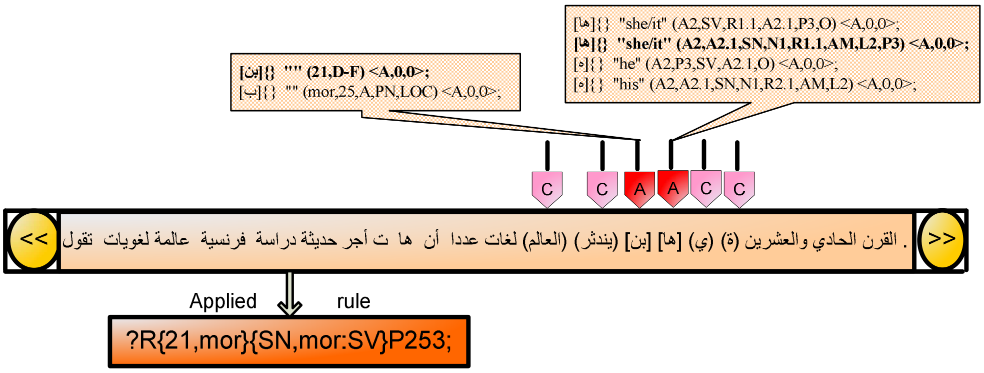

[ها]{} "she/it" (A2,SV,R1.1,A2.1,P3,O) <A,0,0>;
[ها]{} "she/it" (A2,A2.1,SN,N1,R1.1,AM,L2,P3) <A,0,0>;
[ه]{} "he" (A2,P3,SV,A2.1,O) <A,0,0>;
[ه]{} "his" (A2,A2.1,SN,N1,R2.1,AM,L2) <A,0,0>;
[بن]{} "" (21,D-F) <A,0,0>;
[ب]{} "" (mor,25,A,PN,LOC) <A,0,0>;
C C A A C C
<<
>>
Applied rule
?{D-F,^blk,mor}{O,mor}P253;

Figure (39)

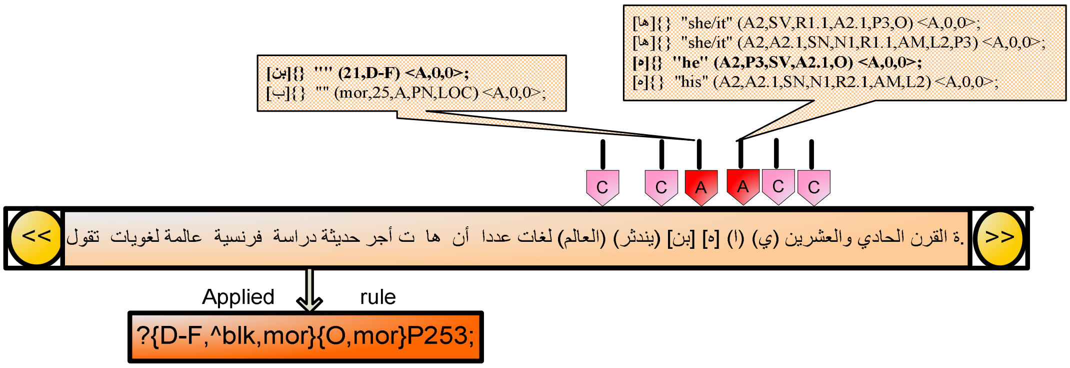
However, this is not correct because this verb is an unaugmented defective verb with final vowel deletion. This means the object pronoun “ها” is not expected to follow; the stem “بنا” should be used instead of “بنى”. Therefore, the backtrack rule in figure (39) is applied to find another solution for the right node suitable for the left node.



[بن]{} "" (21,D-F) <A,0,0>;
[ب]{} "" (mor,25,A,PN,LOC) <A,0,0>;
[ها]{} "she/it" (A2,SV,R1.1,A2.1,P3,O) <A,0,0>;
[ها]{} "she/it" (A2,A2.1,SN,N1,R1.1,AM,L2,P3) <A,0,0>;
[ه]{} "he" (A2,P3,SV,A2.1,O) <A,0,0>;
[ه]{} "his" (A2,A2.1,SN,N1,R2.1,AM,L2) <A,0,0>;
C
C
A
A
C
C
<<
. القرن الحادي والعشرين (ة) (ي) [ها] [بن] (يندثر) (العالم) لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
?R{21,mor}{SN,mor:SV}P253;

Figure (40)

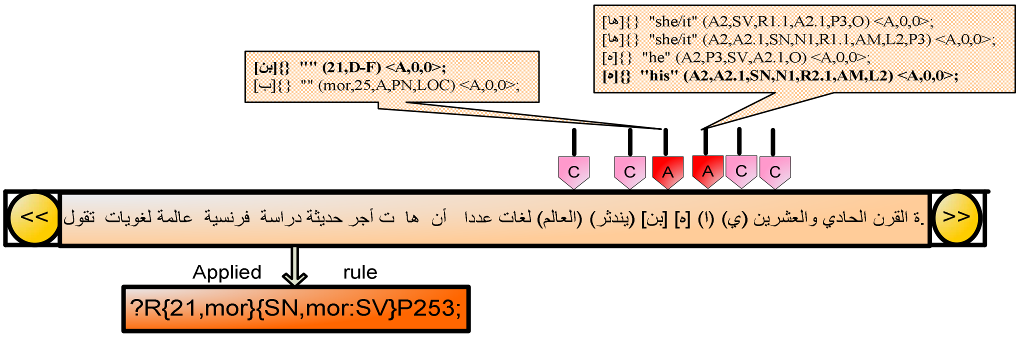
However, the second possibility selected for "ها" is a nominal suffix which is not appropriate to fill this slot (figure (40)).



[بن]{} "" (21,D-F) <A,0,0>;
[ب]{} "" (mor,25,A,PN,LOC) <A,0,0>;
[ها]{} "she/it" (A2,SV,R1.1,A2.1,P3,O) <A,0,0>;
[ها]{} "she/it" (A2,A2.1,SN,N1,R1.1,AM,L2,P3) <A,0,0>;
[ه]{} "he" (A2,P3,SV,A2.1,O) <A,0,0>;
[ه]{} "his" (A2,A2.1,SN,N1,R2.1,AM,L2) <A,0,0>;
C
C
A
A
C
C
<<
ة القرن الحادي والعشرين (ي) (ا) [ه] [بن] (يندثر) (العالم) لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول.
>>
Applied rule
?{D-F,^blk,mor}{O,mor}P253;

Figure (41)

Another backtrack rule rejected this selection. As there is no other “ها” in the dictionary, the EnConverter starts to divide this node into smaller nodes to find a match. Therefore the next possibility is the object pronoun "هـ" (figure (41)).



[بن]{} "" (21,D-F) <A,0,0>;
[ب]{} "" (mor,25,A,PN,LOC) <A,0,0>;
[ها]{} "she/it" (A2,SV,R1.1,A2.1,P3,O) <A,0,0>;
[ها]{} "she/it" (A2,A2.1,SN,N1,R1.1,AM,L2,P3) <A,0,0>;
[ه]{} "he" (A2,P3,SV,A2.1,O) <A,0,0>;
[ه]{} "his" (A2,A2.1,SN,N1,R2.1,AM,L2) <A,0,0>;
C
C
A
A
C
C
<<
ة القرن الحادي والعشرين (ي) (ا) [ه] [بن] (يندثر) (العالم) لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول.
>>
Applied rule
?R{21,mor}{SN,mor:SV}P253;

Figure (42)

Again, the EnConverter rejects the right node and backtracks for another solution (figure (42)). The last possibility available is the possessive nominal suffix which is normally rejected as there is a verb that precedes.

Having tried all the possibilities of the right node without any positive result, the EnConverter shifts backtracking options to the left node and tries to find another solution. As at this point, there is no other solution for the left node, the EnConverter tries to divide it into other smaller nodes. Therefore, the possibility of "ب" to instantiate the left node is under investigation as in figure (43).

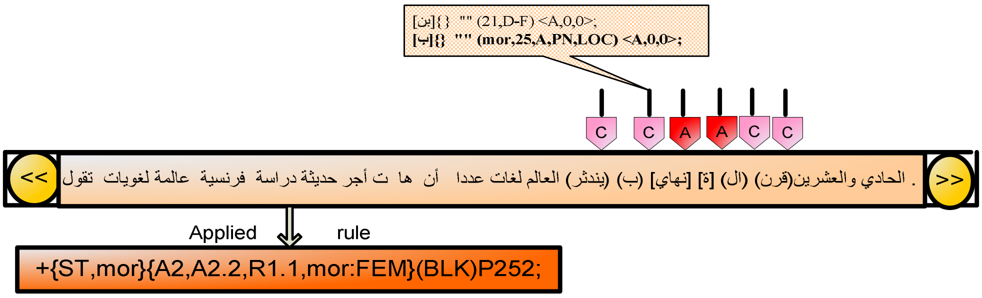

[بن]{ } "" (21,D-F) <A,0,0>;
[ب]{} "" (mor,25,A,PN,LOC) <A,0,0>;
C
C
A
A
C
C
<<
. الحادي والعشرين(قرن) (ال) [ة] [نهاي] (ب) (بندر) العالم لغات عدد أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
+{ST,mor}{A2,A2.2,R1.1,mor:FEM}(BLK)P252;

Figure (43)

With this semi-success in applying rules, the Analysis Windows move one step right which results in a suitable solution for the word "بنهاية" as a noun followed by a nominal feminine suffix. With the application of the composition rule in figure (43), the stem in the right node is given the feature FEM because of the suffix "ة".

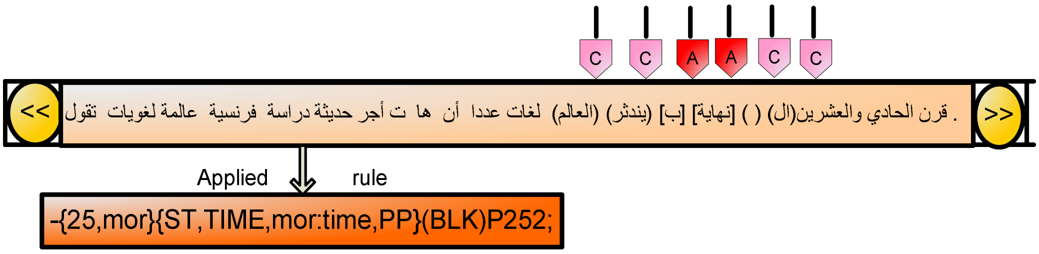

C
C
A
A
C
C
<<
. قرن الحادي والعشرين(ال) ( ) [نهاية] [ب] (بندر) (العالم) لغات عدد أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
-{25,mor}{ST,TIME,mor:time,PP}(BLK)P252;

Figure (44)

As a result of the composition that occurred in figure (43), the EnConverter moves one step backward to find out whether or not another rule can be applied on the new composed nodes (figure (44)). As the left node is a preposition and the right window is a noun, they will be composed together by means of the rule in figure (44), assigning the feature "time" to the stem in the right node. In addition, it marks the right node as starting a prepositional Phrase (PP).

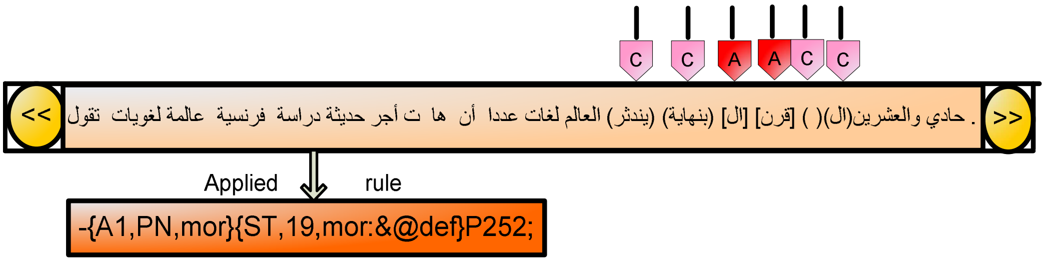

C
C
A
A
C
C
<<
. حادي والعشرين(ال)( ) [قرن] [ال] (بنهاية) (بندر) العالم لغات عدد أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
-{A1,PN,mor}{ST,19,mor:&@def}P252;

Figure (45)

The analysis windows are moving right detecting an article in the left node followed by a noun in the right node. As shown in figure (45), the two nodes are composed together and the UNL attribute "@def" is given to the right node.

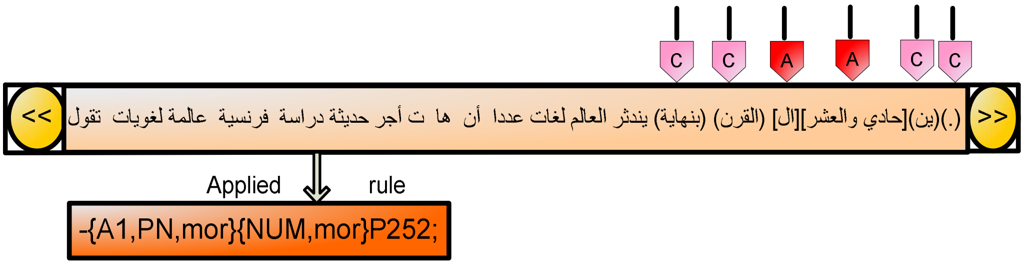
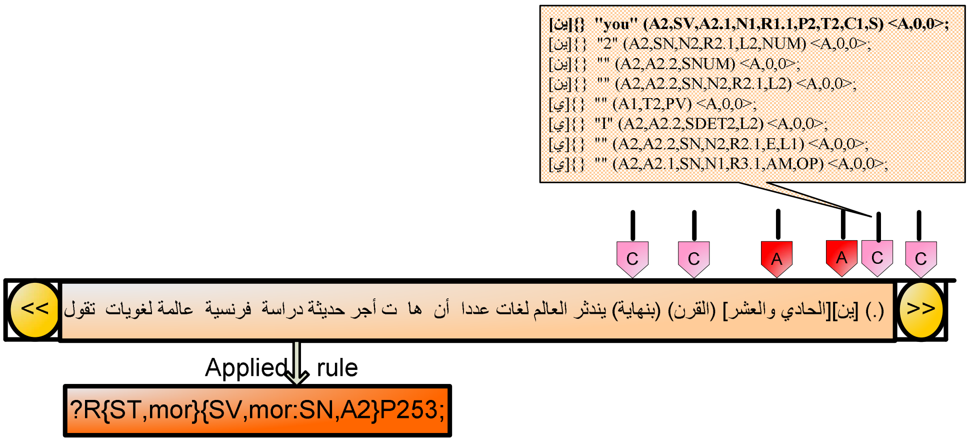

C
C
A
A
C
C
<<
(.)(ين)[حادي والعشر][ال] (القرن) (بنهاية) بندر العالم لغات عدد أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
-{A1,PN,mor}{NUM,mor}P252;

Figure (46)

The next morphemes are the definite article and the Numeral noun on which the composition rule is applied in figure (46) without giving the stem "@def" because of its part of speech.



```
[ين]{}  "you" (A2,SV,A2.1,N1,R1.1,P2,T2,C1,S) <A,0,0>;
[ين]{}  "2" (A2,SN,N2,R2.1,L2,NUM) <A,0,0>;
[ين]{}  "" (A2,A2.2,SNUM) <A,0,0>;
[ين]{}  "" (A2,A2.2,SN,N2,R2.1,L2) <A,0,0>;
[ي]{}  "" (A1,T2,PV) <A,0,0>;
[ي]{}  "I" (A2,A2.2,SDET2,L2) <A,0,0>;
[ي]{}  "" (A2,A2.2,SN,N2,R2.1,E,L1) <A,0,0>;
[ي]{}  "" (A2,A2.1,SN,N1,R3.1,AM,OP) <A,0,0>;
```

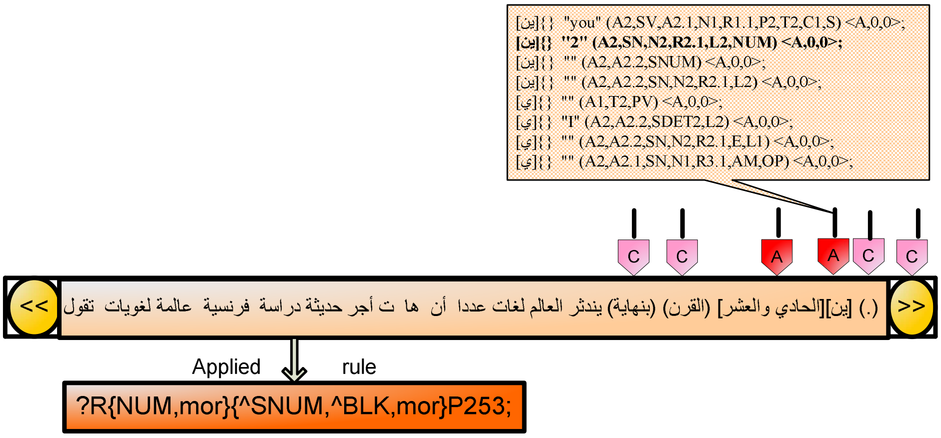
(.) [ين][الحادي والعشر] (القرن) (بنهاية) يندثر العالم لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول

Applied rule

?R{ST,mor}{SV,mor:SN,A2}P253;

Figure (47)

The next morpheme under investigation is the suffix of the Numeral Noun. The first interpretation is a verbal feminine suffix which is rejected by means of the backtrack rule in figure (47) as the previous rule is not a verb.



```
[ين]{}  "you" (A2,SV,A2.1,N1,R1.1,P2,T2,C1,S) <A,0,0>;
[ين]{}  "2" (A2,SN,N2,R2.1,L2,NUM) <A,0,0>;
[ين]{}  "" (A2,A2.2,SNUM) <A,0,0>;
[ين]{}  "" (A2,A2.2,SN,N2,R2.1,L2) <A,0,0>;
[ي]{}  "" (A1,T2,PV) <A,0,0>;
[ي]{}  "I" (A2,A2.2,SDET2,L2) <A,0,0>;
[ي]{}  "" (A2,A2.2,SN,N2,R2.1,E,L1) <A,0,0>;
[ي]{}  "" (A2,A2.1,SN,N1,R3.1,AM,OP) <A,0,0>;
```

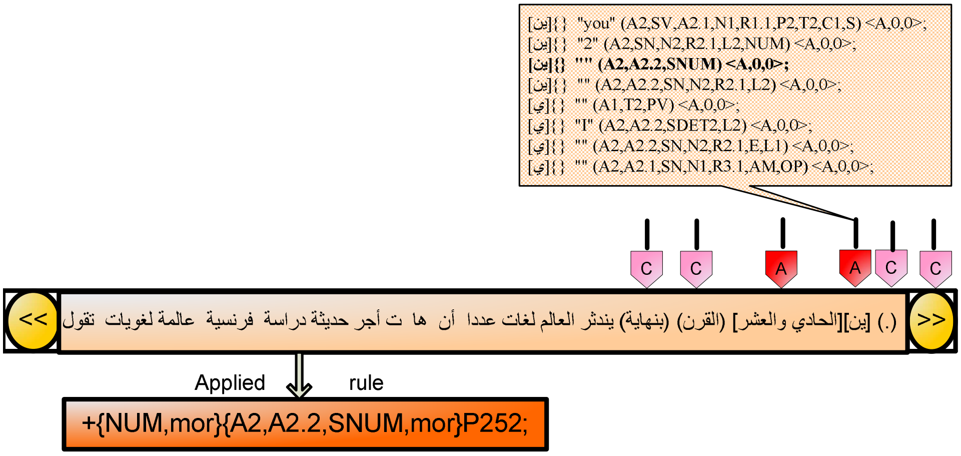
(.) [ين][الحادي والعشر] (القرن) (بنهاية) يندثر العالم لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول

Applied rule

?R{NUM,mor}{^SNUM,^BLK,mor}P253;

Figure (48)

Again the right node is rejected as it is not a case morpheme for Numeral noun of the type in the left node.



```
[ين]{}  "you" (A2,SV,A2.1,N1,R1.1,P2,T2,C1,S) <A,0,0>;
[ين]{}  "2" (A2,SN,N2,R2.1,L2,NUM) <A,0,0>;
[ين]{}  "" (A2,A2.2,SNUM) <A,0,0>;
[ين]{}  "" (A2,A2.2,SN,N2,R2.1,L2) <A,0,0>;
[ي]{}  "" (A1,T2,PV) <A,0,0>;
[ي]{}  "I" (A2,A2.2,SDET2,L2) <A,0,0>;
[ي]{}  "" (A2,A2.2,SN,N2,R2.1,E,L1) <A,0,0>;
[ي]{}  "" (A2,A2.1,SN,N1,R3.1,AM,OP) <A,0,0>;
```

(.) [ين][الحادي والعشر] (القرن) (بنهاية) يندثر العالم لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول

Applied rule

+{NUM,mor}{A2,A2.2,SNUM,mor}P252;

Figure (49)

Finally, the rule in figure (49) could compose the case morpheme to its Numeral stem as the right and left nodes are compatible to each other.

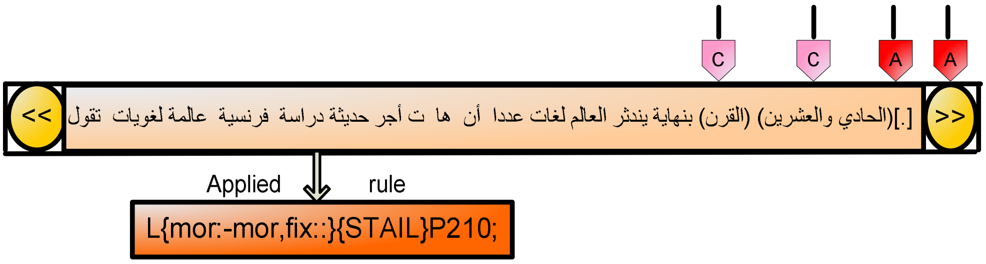

A
A
C
C
<<
[.](الحادي والعشرين) (القرن) بنهاية يندثر العالم لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
L{mor:-mor,fix::}{STAIL}P210;

Figure (50)

The final move of the EnConverter is in progress where it defines a final punctuation mark followed by the STAIL (figure (50)). Having detected the STAIL, the EnConverter realizes that the end of the sentence has been reached. The following is the output of the morphological analysis of the sentence described in section 5.1.4:

```
/<</تقول/عالمة لغويات/فرنسية/دراسة/حديثة/أجر/ت/ها/أن/عددا/لغات/العالم/يندثر/نهاية/القرن/الحادي و العشرين/>>/./

[<<]{} "" (SHEAD) <.,0,0>;.[]{}

[تقول]{}"say(agt>thing,obj>thing)" (mor,blk,&@present,TA,NFAGT,21,ANAGT,OBJ,SAY) <A,0,0>;.[]{}
|-[ت]{}  "" (mor,A1,T2,PV,F) <A,0,0>;.[]{}
|-[قول]{}  "say(agt>thing,obj>thing)" (mor,21,ANAGT,OBJ,SAY) <A,0,0>;.[]{}

[عالمة لغويات]{}  "linguist(icl>person)" (mor,blk,ST,19,AN,FEM) <A,0,0>;.[]{}

[فرنسية] {}"French(aoj>thing)" (mor,blk,FEM,ST,22) <A,0,0>;.[]{}
|-[فرنسي] {} "French(aoj>thing)" (mor,ST,22) <A,0,0>;.[]{}
|-[ة]{} "" (mor,A2,A2.2,SN,N1,R1.1,L2) <A,0,0>;.[]{}

[دراسة ]  {}"study(icl>activity)" (mor,blk,FEM,PP,loc,ST,19,VP) <A,0,0>;.[]{}
|-[دراس ]  {}"study(icl>activity)" (PP,loc,mor,ST,19,VP) <A,0,0>;.[]{}
|-[ة]{} "" (mor,A2,A2.2,SN,N1,R1.1,L2) <A,0,0>;.[]{}

[حديثة] {} "recent(aoj>thing)" (mor,blk,FEM,ST,22) <A,0,0>;.[]{}
|-[حديث] [{}  "recent(aoj>thing)" (mor,ST,22) <A,0,0>;.[]{}
|-[ة] {}"" (mor,A2,A2.2,SN,N1,R1.1,L2) <A,0,0>;.[]{}

[أجر]{} "conduct(agt>thing,obj>thing)" (mor,&@past,TA,21,D-F,ANAGT,OBJ)
<A,0,0>;.[]{}
| -{T1,mor:::}{21,AUG,^&@past,^TA,mor:&@past,TA::}P252;

|-[أ]{}  "" (mor,A1,T1,PV) <A,0,0>;.[]{}
|-[جر]{}  "conduct(agt>thing,obj>thing)" (mor,21,D-F,ANAGT,OBJ) <A,0,0>;.[]{}

[ت]{}  "she/it" (mor,A2,SV,A2.1,N1,R1.1,P3,T1,S,F) <A,0,0>;.[]{}

[ها]{}  "she/it" (mor,blk,A2,SV,R1.1,A2.1,P3,O) <A,0,0>;.[]{}

[أن]{}  "" (mor,blk,DET2) <A,0,0>;.[]{}

[عددا]{}  "number(qua<thing)" (mor,blk,ST,QUA,19) <A,0,0>;.[]{}
|-[عدد]{}  "number(qua<thing)" (mor,ST,QUA,19) <A,0,0>;.[]{}
|-[ا]{}  "" (mor,A2,A2.2,SN,N1,C) <A,0,0>;.[]{}

[لغات] {} "language(icl>science)" (mor,blk,&@pl,FEM,PP,ST,19) <A,0,0>;.[]{}
|-[لغ]   {}"language(icl>science)" (PP,mor,ST,19) <A,0,0>;.[]{}
|-[ات]{}  "" (mor,A2,A2.2,SN,N2,R1.1,L1) <A,0,0>;.[]{}

[العالم]{}  "world(icl>region)" (mor,blk,&@def,ST,19,AN) <A,0,0>;.[]{}
```

```
|-[ال]{}  "" (mor,A1,PN) <A,0,0>;.[]{}
|-[عالم]{}  "world(icl>region)" (mor,ST,19,AN) <A,0,0>;.[]{}

[يندثر]{}  "disappear(obj>thing)" (mor,blk,&@future,TA,21,AUG,OBJ) <A,0,0>;.[]{}
|-[ي]{}  "" (mor,A1,T2,PV) <A,0,0>;.[]{}
|-[ندثر]{}  "disappear(obj>thing)" (mor,21,AUG,OBJ) <A,0,0>;.[]{}

[نهاية]{}  "end(icl>time)" (mor,blk,FEM,PP,loc,ST,19,TIME) <A,0,0>;.[]{}
|-[نهاي]{}  "end(icl>time)" (PP,loc,mor,ST,19,TIME) <A,0,0>;.[]{}
|-[ة]{}  "" (mor,A2,A2.2,SN,N1,R1.1,L2) <A,0,0>;.[]{}

[القرن] {}"century(icl>time)" (mor,blk,&@def,ST,19) <A,0,0>;.[]{}
|-[ال] {} "" (mor,A1,PN) <A,0,0>;.[]{}
|-[قرن] {}"century(icl>time)" (mor,ST,19) <A,0,0>;.[]{}

[21"  {}[الحادي والعشرين" (mor,ST,NUM,22) <A,0,0>;.[]{}
| +{NUM,mor:::}{A2,A2.2,N2,mor:::}P252;
|-[21"  {}[الحادي والعشر" (mor,ST,NUM,22) <A,0,0>;.[]{}
| | -{A1,PN,mor:::}{NUM,mor:::}P252;
| |-[ال]{}  "" (mor,A1,PN) <A,0,0>;.[]{}
| |-[21"  {}[حادي والعشر" (mor,ST,NUM,22) <A,0,0>;.[]{}
|-[ين]{}  "" (mor,A2,A2.2,SN,N2,R2.1,L2) <A,0,0>;.[]{}

[.]{}  "" (mor,PUNCT,FS) <A,0,0>;.[]{}

[>>]{}  "" (STAIL) <.,0,0>;.[]{}
```

Based on the morphological analysis (segmentation and identification of the nodes of the sentence), Universal words below are generated to be ready to receive semantic relations in the next stage of the grammar:

```
say(agt>thing,obj>thing):01.@present
linguist(icl>person):05
French(aoj>thing):0I
study(icl>activity):0S
recent(aoj>thing):0Y
conduct(agt>thing,obj>thing):15.@past
she/it:17
she/it:18
number of(qua<thing):1E
language(icl>science):1M.@pl
world(icl>region):1T.@def
disappear(obj>thing):23.@future
end(icl>time):29
century(icl>time):2H.@def
21:2N
".":30
```

After finishing the morphological analysis, all nodes in the sentence will be marked in a way that will not make them targets for morphological rules anymore. A left shift rule is applied from left to right to mark every node with the feature “fix” to denote the end of morphological analysis (figure (51)). Finishing marking nodes, the analysis windows will reach the beginning of the sentence and start building semantic relations between the Universal Words generated in the previous stage.

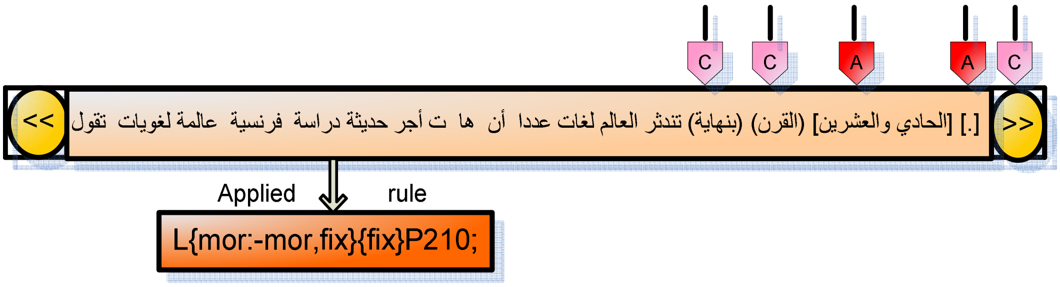

C
C
A
A
C
<<
[.] [الحادي والعشرين] (القرن) (بنهاية) تندثر العالم لغات عددا أن ها ت أجر حديثة دراسة فرنسية عالمة لغويات تقول
>>
Applied rule
L{mor:-mor,fix}{fix}P210;

Figure (51)

5.1.4 Building semantic relations (Hyper Semantic Network):

After the morphological analysis stage has been accomplished and UWs have been extracted, the stage of assigning relations starts. At this point it is very important to clarify the following two points. First, relations in the UNL system are binary; this means that the relation stage aims at building relations between every pair of UWs. Second, constructing semantic relations is a shallow form of semantic representation which is a case-role analysis, which identifies roles such as agent, patient, source, time, purpose, and destination. Automatic, accurate and wide-coverage techniques that can annotate naturally occurring text with semantic roles and relations can play a key role in NLP applications such as Information Extraction, Question Answering and Summarization. Shallow semantic parsing – the process of assigning a simple WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW, etc. structure to sentences in text, is the process of producing such a markup (Johnson (1988)).

Every sentence contains a main predicate or a main concept that the speaker focuses on. Every predicate in turn has a number of arguments which completes the meaning of the predicate. Every argument may have modifiers that complete/elaborate the meaning of the sentences or phrases. Arguments can be classified into two types: primary and secondary. Primary arguments are those arguments that are necessary to complete the meaning of the verb, the main predicate of the sentence. This primary argument in some sentences is not mentioned but we understand it from the whole sentence. Examples of primary arguments are: agent, patient and goal. Secondary arguments are not necessary for completing the meaning of verb. They are modifiers that are different from one sentence to another. Examples of secondary arguments are: manner, reason, and purpose.

According to what is mentioned above, our design of the relation stage is divided into two sub-stages. The first sub-stage aims at constructing relations between UWs representing modifiers while the second sub-stage deals with constructing relations between UWs representing the main skeleton of the sentence. This section will continue discussing in some more details how the UWs generated from the sentence discussed in section 5.1.3.1 are linked together till a UNL hyper semantic network is generated.

5.1.4.1 A corpus-based example of formalizing semantic relation:

In figure (51), trace is stopped by the end of segmenting and identifying morphemes of the sentence under analysis. By the end of applications of left shift rules, the focus of the EnConverter is given to the beginning of the processing (SHEAD) and the first node in the node list (figure (52)).

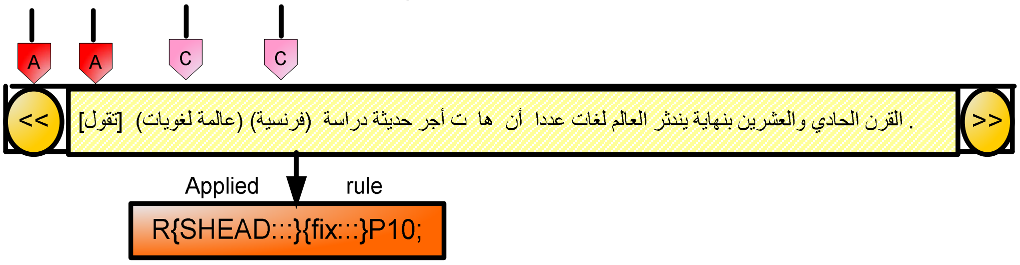

A
A
C
C
<<
. القرن الحادي والعشرين بنهاية يندثر العالم لغات عددا أن ها ت أجر حديثة دراسة (فرنسية) (عالمة لغويات) [تقول]
>>
Applied rule
R{SHEAD:::}{fix:::}P10;

Figure (52)

As the system is still on the first node, it could be normal to understand that no rule can apply. Therefore, the grammar moves one step right (figure (53)).

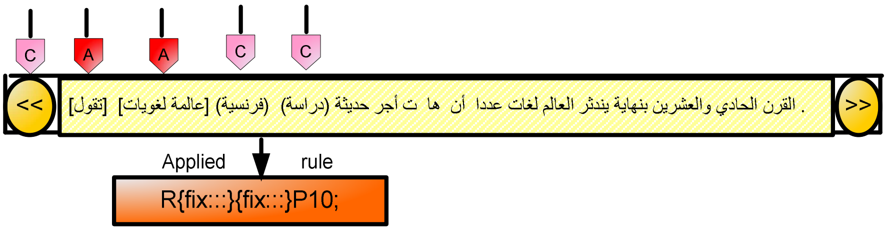

C
A
A
C
C
<<
القرن الحادي والعشرين بنهاية ينذر العالم لغات عددا أن ها ت أجر حديثة (دراسة) (فرنسية) [عالمة لغويات] [تقول] .
>>
Applied
rule
R{fix:::}{fix:::}P10;

Figure (53)

After the right shift rule has been applied, the left Analysis Window is on the node (تقول) and the right analysis window is on (عالمة لغويات), again no rule applies because the first right condition window is an adjective which might be considered as a modifier to the right node. Therefore, the right shift rule applies to move analysis windows to the next node in order to give priority to link modifiers to their heads first (figure (53)).

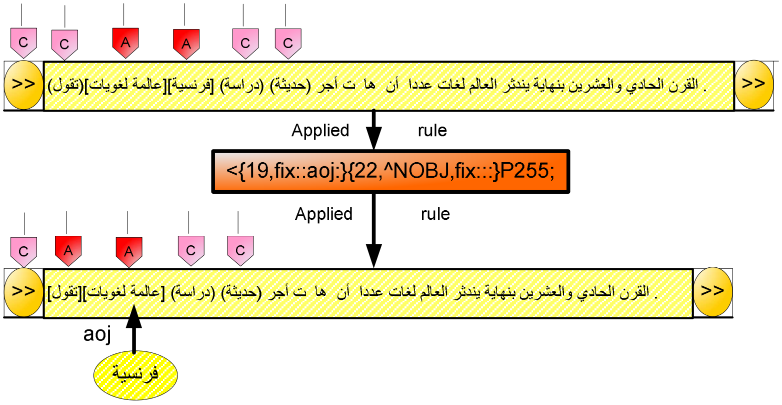

C
C
A
A
C
C
>>
القرن الحادي والعشرين بنهاية ينذر العالم لغات عددا أن ها ت أجر (حديثة) (دراسة) [فرنسية][عالمة لغويات](تقول) .
>>
Applied
rule
<{19,fix::aoj:}{22,^NOBJ,fix:::}P255;
Applied
rule
C
A
A
C
C
>>
القرن الحادي والعشرين بنهاية ينذر العالم لغات عددا أن ها ت أجر (حديثة) (دراسة) [عالمة لغويات][تقول] .
>>
aoj
فرنسية

Figure (54)

As the left analysis window is on (عالمة لغويات) and the right analysis window is on (فرنسية), and the dictionary supplies information about the parts of speech of the two nodes, the rule in figure (54) applies to hold an “aoj” (means ‘thing with attribute’) relation between the two concepts. As a result of applying the rule, the right node leaves the node list to the node net. It is very important at this point to note the direction of semantic relations in UNL. In the current case, the right node assigns the relation while the left node receives this relation. After application of the “aoj” relation, the Analysis Windows move back one step to test whether or not another rule can apply on the ‘new’ list of nodes, after the departure of (فرنسية) from the node list.

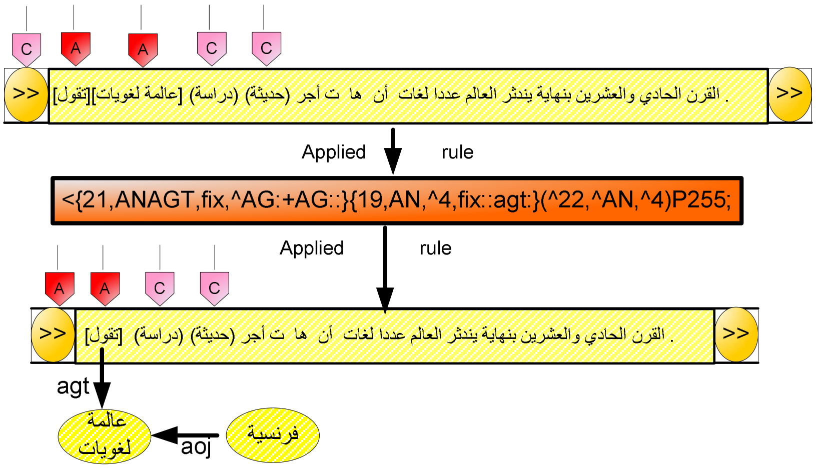

C
A
A
C
C
>>
القرن الحادي والعشرين بنهاية ينذر العالم لغات عددا أن ها ت أجر (حديثة) (دراسة) [عالمة لغويات][تقول] .
>>
Applied
rule
<{21,ANAGT,fix,^AG:+AG::}{19,AN,^4,fix::agt:}(^22,^AN,^4)P255;
Applied
rule
A
A
C
C
>>
القرن الحادي والعشرين بنهاية ينذر العالم لغات عددا أن ها ت أجر (حديثة) (دراسة) [تقول] .
>>
agt
عالمة لغويات
aoj
فرنسية

Figure (55)

Moving back, the left analysis window is on (تقول) while the right analysis window is on (عالمة لغويات), an “agt” relation applies between the two nodes making the left node, the verb, assigns the relation and the right node receives it (figure (55)). Accordingly, after the departure of (عالمة لغويات) from the node list, the analysis windows move back to examine the new situation.

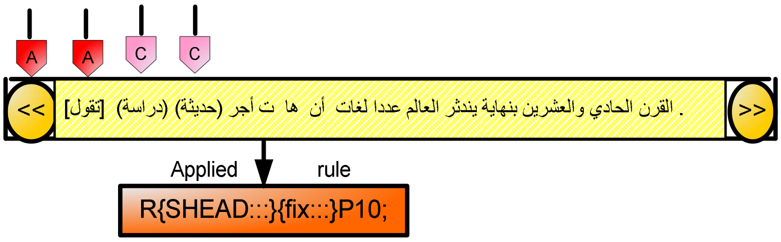

A
A
C
C
<<
. القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات أن ها ت أجر (حديثة) (دراسة) [تقول]
>>
Applied rule
R{SHEAD:::}{fix:::}P10;

Figure (56)

Returning back, the EnConverter discovers that the start of the processing (SHEAD) has been reached again, therefore, a right shift rule applies to move the flow of processing to next node (figure (56)), as happened before in figure (52).

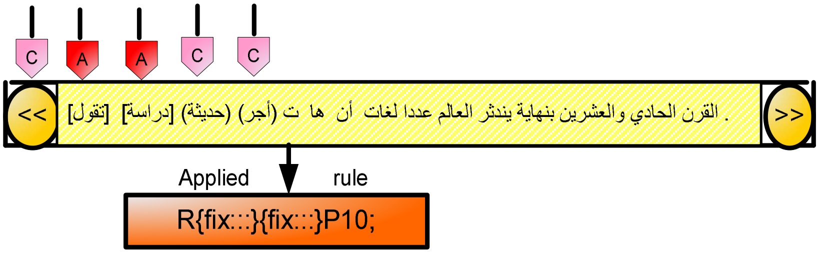

C
A
A
C
C
<<
. القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات أن ها ت (أجر) (حديثة) [دراسة] [تقول]
>>
Applied rule
R{fix:::}{fix:::}P10;

Figure (57)

After the right shift, the left analysis window is on (تقول) and the right analysis window is on (دراسة) no rule applies because the first right condition window is an adjective which might be consider as a modifier to the right node. Therefore the right shift rule applies to move to the next node to give priority to link modifiers to their heads first (figure (57)).

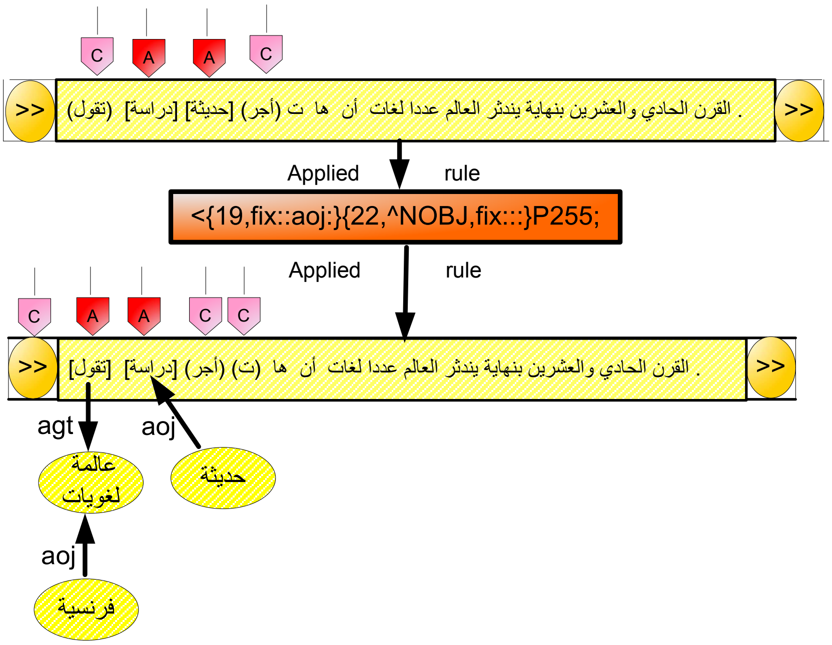

C
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات أن ها ت (أجر) [حديثة] [دراسة] (تقول)
>>
Applied rule
<{19,fix::aoj:}{22,^NOBJ,fix:::}P255;
Applied rule
C
A
A
C
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات أن ها (ت) (أجر) [دراسة] [تقول]
>>
agt
aoj
عالمة لغويات
حديثة
aoj
فرنسية

Figure (58)

As the left analysis window is on (دراسة) and the right analysis window is on (حديثة), and the dictionary supplies information about the parts of speech of the two nodes, the rule in figure (58) applies to hold an 'aoj' relation between the two concepts as the two nodes agrees in number and gender. As a result of applying the rule, the right node leaves the node list to the node net. After application of the "aoj" relation, analysis windows move back one step to test whether or not another rule can apply on the new list of nodes, after the departure of (حديثة) from the node list.

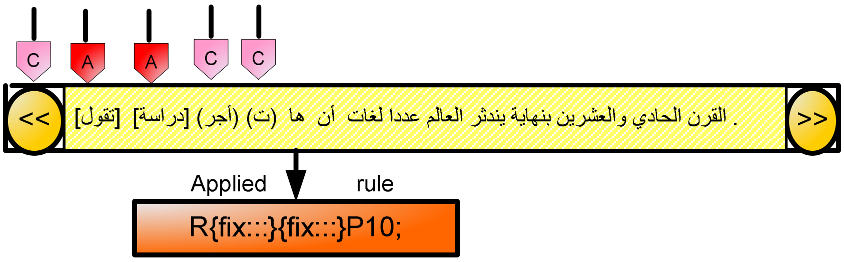

C
A
A
C
C
<<
القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات أن ها (ت) (أجر) [دراسة] [تقول]
>>
Applied
rule
R{fix:::}{fix:::}P10;

Figure (59)

In the new situation in figure (59) although the left Analysis Window is a verb (تقول) and the right window is a noun (دراسة) which might be linked together with a given relation, but this decision is delayed because the first right condition window is a verb and the noun in the right analysis window might be a modifier to this verb. Therefore, the right shift rule applies to test the next node.

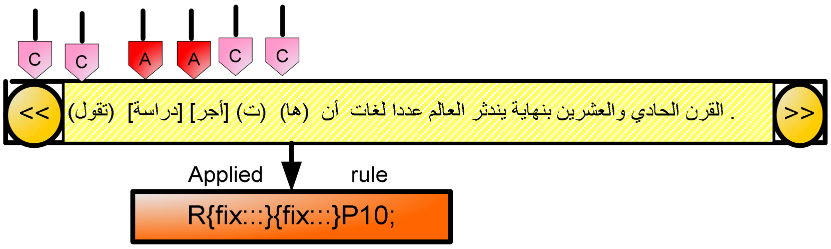

C
C
A
A
C
C
<<
القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات أن (ها) (ت) [أجر] [دراسة] (تقول)
>>
Applied
rule
R{fix:::}{fix:::}P10;

Figure (60)

When the EnConverter stops by (دراسة) and (أجر), it realizes that the right node is a verb. As there is a pronoun which follows the verb, according to the design adopted, the EnConverter tries to construct a relation between the verb and its pronouns first. Consequently, the EnConverter tries to find the agent first as the verb is marked as a 'do' transitive verb in the dictionary. Therefore, the right shift rule in figure (60) works to shift the processing one step right.

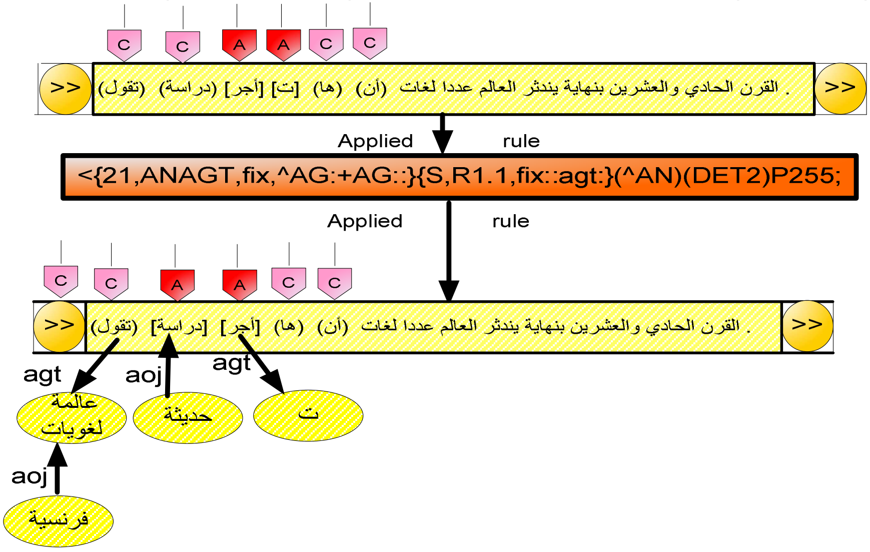

C
C
A
A
C
C
>>
القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات (أن) (ها) [ت] [أجر] (دراسة) (تقول)
>>
Applied
rule
<{21,ANAGT,fix,^AG:+AG::}{S,R1.1,fix::agt:}(^AN)(DET2)P255;
Applied
rule
C
C
A
A
C
C
>>
القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات (أن) (ها) [أجر] [دراسة] (تقول)
>>
agt
aoj
agt
عالمة لغويات
حديثة
ت
aoj
فرنسية

Figure (61)

So far, the left analysis window is on (أجر), the right one is on (ت) and the dictionary supplies information about the parts of speech of the two nodes, the rule in figure (61) applies to hold an 'agt' relation between the two concepts. As a result of applying the rule, the right node leaves the node list to the node net. After application of the "agt" relation, analysis windows move back one step to test whether or not another rule can apply on the new list of nodes, after the departure of (ت) from the node list.

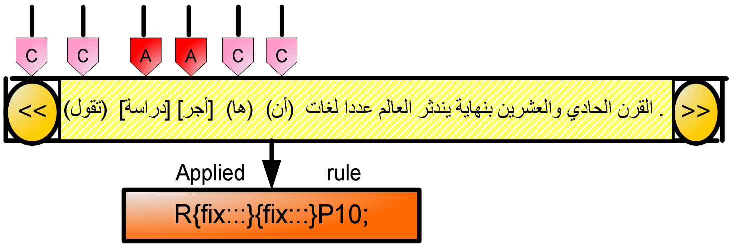

C C A A C C
<<
القرن الحادي والعشرين بنهاية ينذر العالم عددا لغات (أن) (ها) [أجر] [دراسة] (تقول) .
>>
Applied rule
R{fix:::}{fix:::}P10;

Figure (61)

When the EnConverter stops by (دراسة) and (أجر), also no rule can apply because the first right left condition window is a pronoun and the EnConverter gives priority to linking the verb to its pronoun first. Therefore, the right shift rule in figure (62) works to shift the processing one step right.

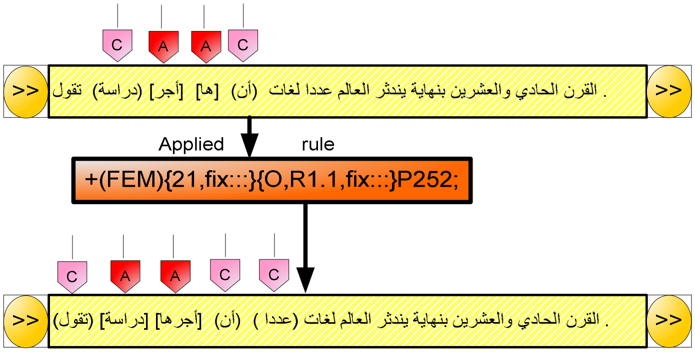

C A A C
>>
القرن الحادي والعشرين بنهاية ينذر العالم عددا لغات (أن) [ها] [أجر] (دراسة) تقول .
>>
Applied rule
+(FEM){21,fix:::}{O,R1.1,fix:::}P252;
C A A C C
>>
القرن الحادي والعشرين بنهاية ينذر العالم لغات (عددا) (أن) [أجرها] [دراسة] (تقول) .
>>

Figure (63)

As for the situation after the movement to the right, the left analysis window is a verb (أجر) which is preceded by a noun that agrees with the pronoun that follows the verb. Therefore, the priority is given to consider the noun in the left condition window as the object of the verb in the left analysis window; the pronoun that follows this verb makes this possibility more probable. This highlights that the relation between the pronoun and the verb is redundant because it is expressed by the noun that precedes the verb. Accordingly, the rule in figure (63) is applied to compose the pronoun to the verb. The analysis windows move back one step to test whether or not there is rule to apply on the new node list.

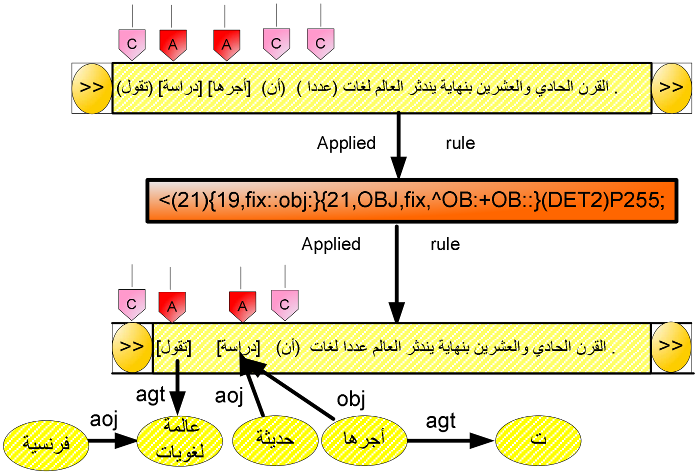

C A A C C
>>
القرن الحادي والعشرين بنهاية ينذر العالم لغات (عددا) (أن) [أجرها] [دراسة] (تقول) .
>>
Applied rule
<(21){19,fix::obj:}{21,OBJ,fix,^OB:+OB::}(DET2)P255;
Applied rule
C A A C
>>
القرن الحادي والعشرين بنهاية ينذر العالم عددا لغات (أن) [دراسة] [تقول]
>>
agt
aoj
obj
aoj
agt
فرنسية
عالمة لغويات
حديثة
أجراها
ت

Figure (64)

Unlike in figure (63), the verb in the right analysis window in figure (64) is preceded by a noun but it is not followed by a pronoun, therefore, a rule is applied to construct an ‘obj’ relation between the left and right nodes in figure (64). After the right node has left the node list, the analysis windows move back to test whether or not there is a rule applicable on the new node list.

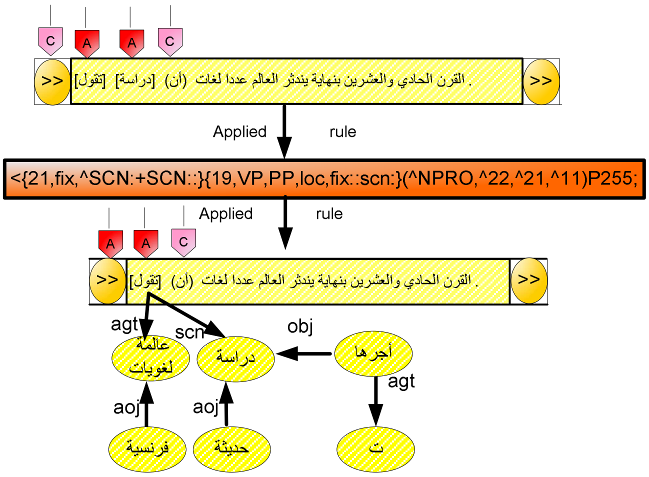

C
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات (أن) [دراسة] [تقول]
>>
Applied
rule
<{21,fix,^SCN:+SCN::}{19,VP,PP,loc,fix::scn:}(^NPRO,^22,^21,^11)P255;
Applied
rule
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم عددا لغات (أن) [تقول]
>>
agt
scn
obj
عالمة لغويات
دراسة
أجرها
agt
aoj
aoj
فرنسية
حديثة
ت

Figure (65)

Now the left analysis window is on (تقول) and the right one is on (دراسة). As the right condition window is on an object marker (أن) for verbs like that in the left analysis window in figure (65), the 'scn' (scene 'virtual world') relation is applied. The node net so far can be seen in figure (65).

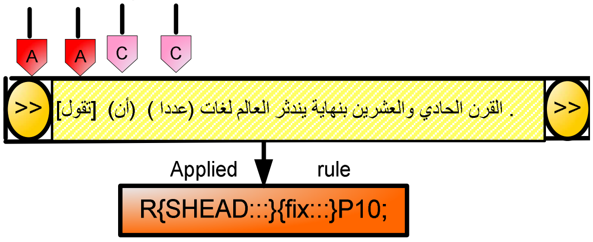

A
A
C
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم لغات (عددا) (أن) [تقول]
>>
Applied
rule
R{SHEAD:::}{fix:::}P10;

Figure (66)

The EnConverter moves left where it discovers that the start of the processing (SHEAD) has been reached again, therefore, a right shift rule applies to move the flow of processing to next node (figure (66)), as happened before in figure (52).

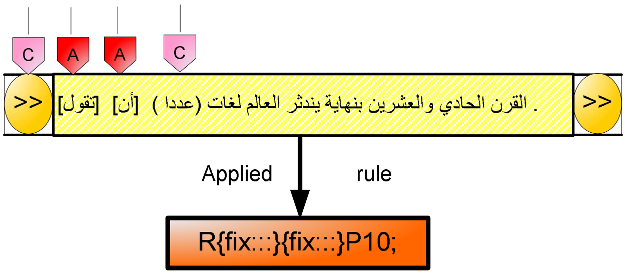

C
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم لغات (عددا) [أن] [تقول]
>>
Applied
rule
R{fix:::}{fix:::}P10;

Figure (67)

In figure (67), no rule applies between the verb (تقول) and its object marker (أن), therefore, a right shift rule is applied to move inside the object clause expected after the object marker.

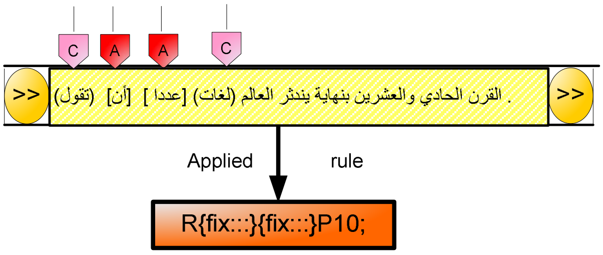

C
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر العالم (لغات) [عددا] [أن] (تقول)
>>
Applied
rule
R{fix:::}{fix:::}P10;

Figure (68)

Still in figure (68), no relation can be held between the object marker and the noun the follows, therefore, a right shift rule applies to examine the next node.

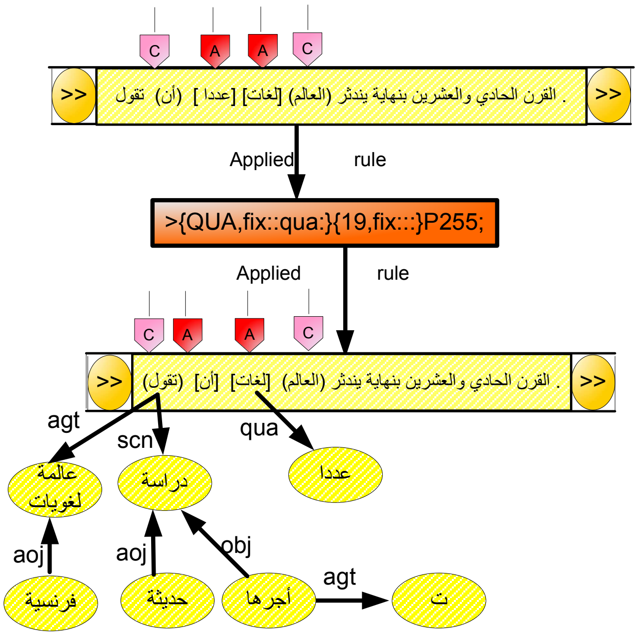

C
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر (العالم) [لغات] [عددا] (أن) تقول
>>
Applied
rule
>{QUA,fix::qua:}{19,fix:::}P255;
Applied
rule
C
A
A
C
>>
. القرن الحادي والعشرين بنهاية يندثر (العالم) [لغات] [أن] (تقول)
>>
agt
scn
qua
عالمة لغويات
دراسة
عددا
aoj
aoj
obj
فرنسية
حديثة
أجرها
agt
ت

Figure (69)

As a result of the right shift the left analysis window is on (عدد) and the right analysis window is on (لغات). The left node is tagged in the dictionary as capable of holding a ‘qua’ (quantity) relation under certain context, therefore a ‘qua’ relation is between the two nodes (figure (69)).

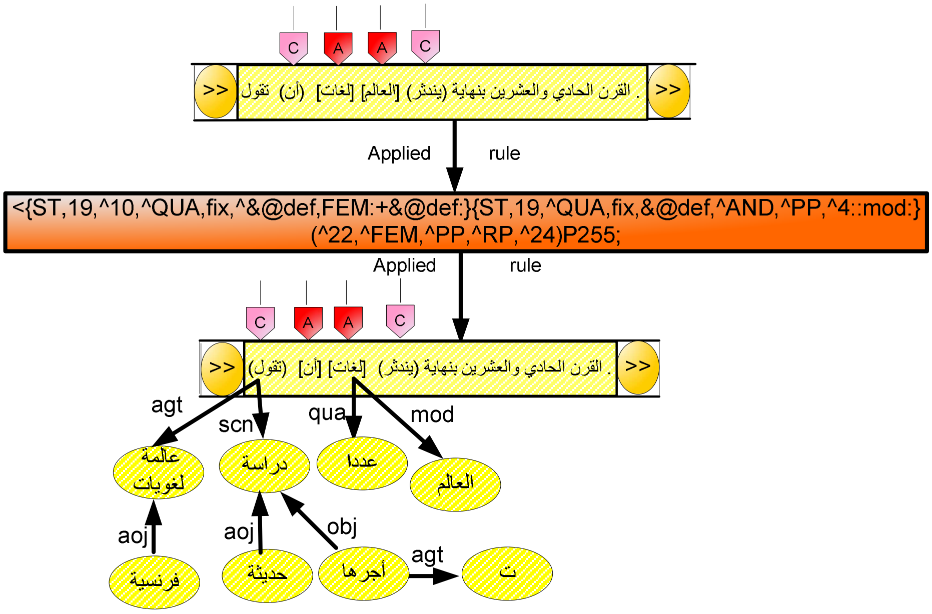

C
A
A
C
>>
. القرن الحادي والعشرين بنهاية (يندثر) [العالم] [لغات] (أن) تقول
>>
Applied
rule
<{ST,19,^10,^QUA,fix,^&@def,FEM:+&@def:}{ST,19,^QUA,fix,&@def,^AND,^PP,^4::mod:}(^22,^FEM,^PP,^RP,^24)P255;
Applied
rule
C
A
A
C
>>
. القرن الحادي والعشرين بنهاية (يندثر) [لغات] [أن] (تقول)
>>
agt
scn
qua
mod
عالمة لغويات
دراسة
عددا
العالم
aoj
aoj
obj
فرنسية
حديثة
أجرها
agt
ت

Figure (70)

As a result of the right shift, the left analysis window is on (لغات) and the right one is on (العالم). The situation of two successive nouns where the first one is indefinite and the second is definite highlights a ‘mod’ relation between them. As the result of applying the rule in figure (70) the right node leaves the node list to the node net, therefore, the analysis windows move back to the left seeking for possible relations. However, nothing applies which leads the EnConverter to move right.

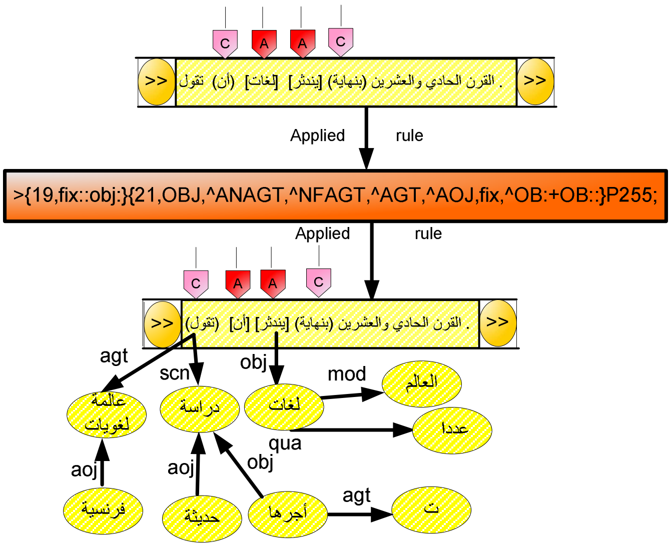

C
A
A
C
>>
. القرن الحادي والعشرين (بنهاية) [يندثر] [لغات] (أن) تقول
>>
Applied
rule
>{19,fix::obj:}{21,OBJ,^ANAGT,^NFAGT,^AGT,^AOJ,fix,^OB:+OB::}P255;
Applied
rule
C
A
A
C
>>
. القرن الحادي والعشرين (بنهاية) [يندثر] [أن] (تقول)
>>
agt
scn
obj
mod
عالمة لغويات
دراسة
لغات
العالم
عددا
qua
aoj
aoj
obj
فرنسية
حديثة
أجرها
agt
ت

Figure (71)

As a result of the right shift, the left analysis window is on (لغات) and the right one is on (يندثر). There is 'obj' relation which applies between them as the right node is a verb which is tagged in the dictionary as an 'occur' verb that does not need an 'agent'. As the result of applying the rule in figure (71) the left node leaves the node list to the node net. Therefore, the analysis windows move back to test the right node with previous node (the one that precedes the left node) (figure (72)).

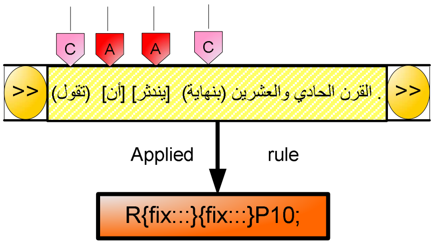


Figure (72)

As a result, the left analysis window is on (أن) and the right one is on (يندثر) and there is no rule to apply, therefore, the right shift rule applies to move the flow of the processing to the next node. Nothing applies, analysis windows continue moving right till the nodes in figure (73) are reached.

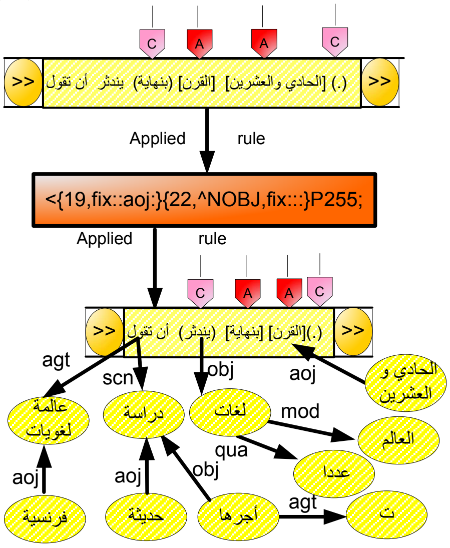
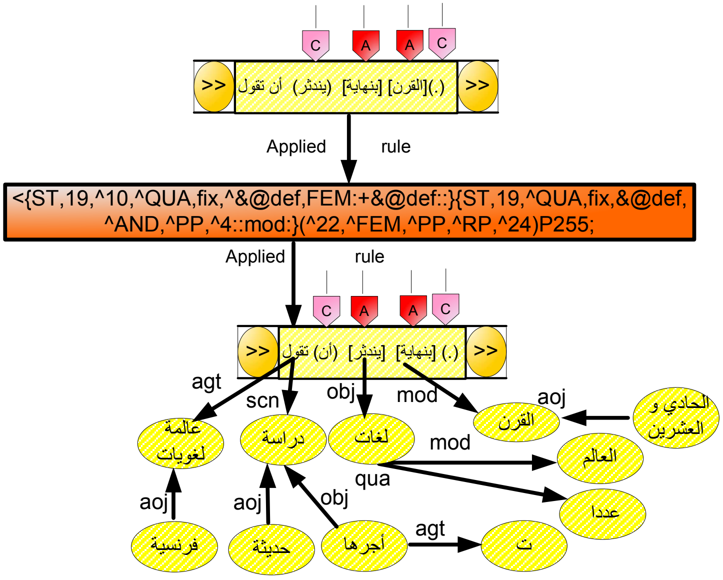


Figure (73)

The ‘aoj’ relation is possible between (القرن) and (الحادي و العشرين) as the left node is a noun followed by an ordinal number. As the result of applying the rule in figure (73) the left node leaves the node list to the node net, giving the possibility to (بنهاية) and (القرن) to be tested against each other.



C
A
A
C
>> أن تقول (يندثر) [بنهاية] [القرن](.) >>
Applied
rule
<{ST,19,^10,^QUA,fix,^&@def,FEM:+&@def::}{ST,19,^QUA,fix,&@def,
^AND,^PP,^4::mod:}(^22,^FEM,^PP,^RP,^24)P255;
Applied
rule
C
A
A
C
>> تقول (أن) [يندثر] [بنهاية] (.) >>
agt
scn
obj
mod
aoj
عالمة لغويات
دراسة
لغات
القرن
الحادي و العشرين
mod
العالم
qua
عددا
aoj
aoj
obj
agt
فرنسية
حديثة
أجرها
ت

Figure (74)

The ‘mod’ relation formally occurs between two nodes when they are successive where the first one is indefinite and the second one is definite. As the result of applying the rule in figure (74) (the right modification rule), the left node leaves the node list to the node net. As happens after the application of rules, analysis windows move back one step to test the possibility of the application of any rule that has been previously inapplicable.

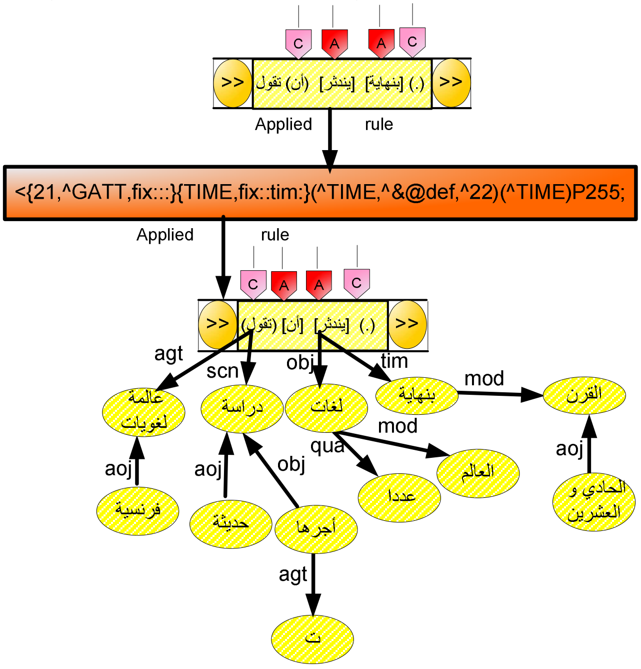


Figure (75)

When the analysis windows are on (يندثر) and (بنهاية), ‘tim’ relation formally occurs between them because the left node is a verb while the right node has the feature ‘TIME’ given the condition that this right node is not followed by neither an adjective nor any node that has the feature ‘TIME’. As a result of applying the rule in figure (75) the left node goes to the node net.

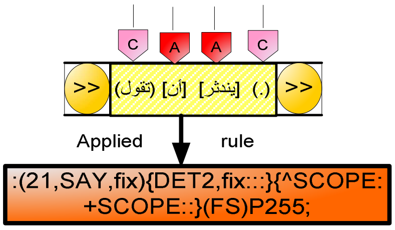


Figure (76)

As a result of the application of the previous rule, the left analysis window is on (أن) while the right one is on (يندثر), and the condition window that follows the right analysis window is a FS (full stop). The FS marks the end of the object clause of the say verb (تقول). Accordingly, the right analysis window is tagged with the feature ‘SCOPE’ which will be used later to hold an ‘object’ relation between the whole clause and the say verb (figure (76)). Based on the application of the rule of the attribute changing in figure (76), the analysis windows move left and stops at (تقول) and (أن) (figure (77)). At this situation, a deletion rule is applied to delete the object marker as it is preceded by a say verb and followed by a SCOPE. In addition the object marker (أن) is no longer needed as the object clause has been determined. Because of the deletion of the object marker, analysis windows move one step left (figure (77)).

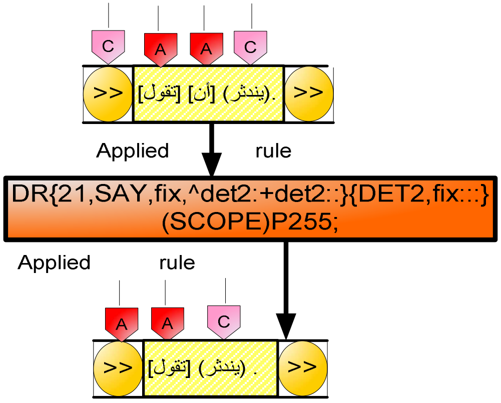


Figure (77)

As the EnConverter reaches the SHEAD, nothing applies. Therefore, a right shift rule works to shift the analysis windows to the right to find applicable rules.

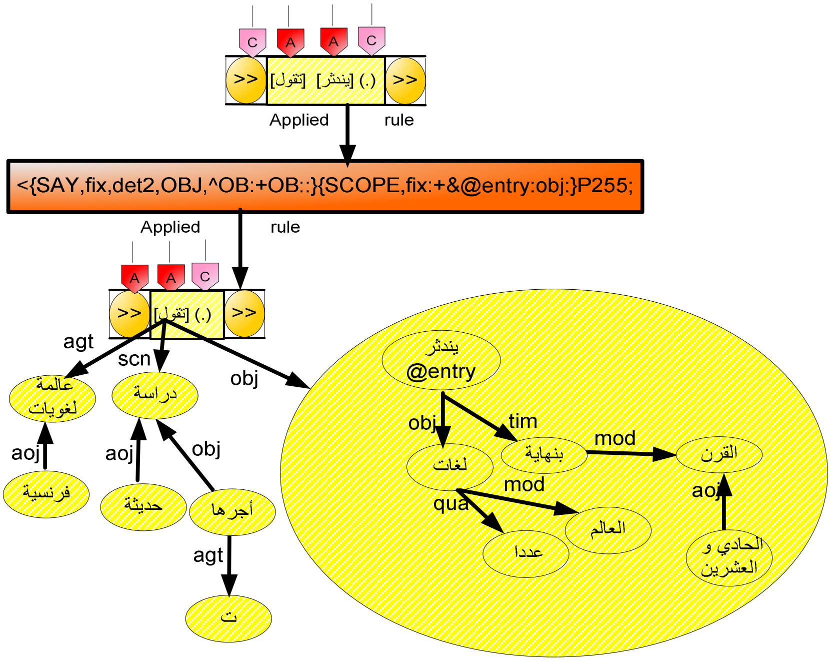


Figure (78)

As a result of right shift the left analysis window is on (تقول) and the right one is on (يندثر). The rule in figure (78) is applied making an ‘obj’ relation between the left node and the scope headed by the node (يندثر).

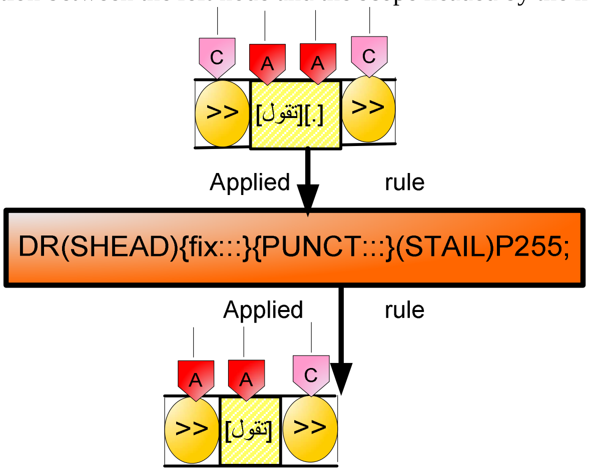


Figure (79)

The semi-final situation is that the left analysis window is on (تقول) and the right one is on a full stop (FS). Therefore the rule in figure (79) is applied to delete the FS. As happens after the application of rules, the analysis windows move left reaching the SHEAD. As no rule in this situation, a right shift rule applies to shift the analysis windows to the right to find applicable rules.

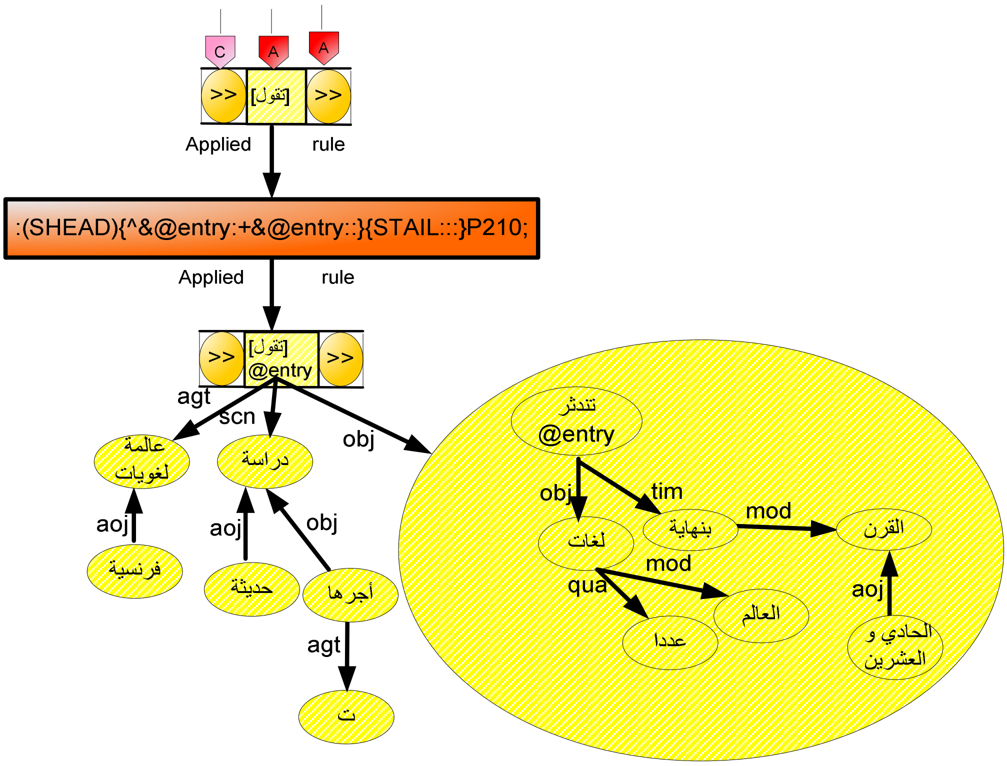


Figure (80)

As a result of the right shift, the left analysis window is on (تقول) and right one is on (STAIL). As the EnConverter reached the final state (the STAIL), the rule in figure (80) is applied to give the last node in the node list the attribute ‘@entry’. The last node remains in the node list is the main predicate of the sentence. Finally, at this situation, the UNL expression is generated as follows:

```
{org}
تقول عالمة لغويات فرنسية في دراسة حديثة أجرتها أن عددا من لغات العالم سوف يندثر بنهاية القرن الحادي والعشرين.
{/org}
{unl}
obj(say(agt>thing,obj>thing):01.@entry.@present,    :01)
scn(say(agt>thing,obj>thing):01.@entry.@present,    study(icl>activity):0S)
agt(say(agt>thing,obj>thing):01.@entry.@present,    linguist(icl>person):05)
aoj(French(aoj>thing):0I,  linguist(icl>person):05)
obj(conduct(agt>thing,obj>thing):15.@past, study(icl>activity):0S)
aoj(recent(aoj>thing):0Y,  study(icl>activity):0S)
agt(conduct(agt>thing,obj>thing):15.@past, she:17)
tim:01(disappear(obj>thing):23.@entry.@future,      end(icl>time):29.@def)
obj:01(disappear(obj>thing):23.@entry.@future,      language(icl>science):1M.@def.@pl)
mod:01(language(icl>science):1M.@def.@pl,           world(icl>region):1T.@def)
qua:01(language(icl>science):1M.@def.@pl,           number of(qua<thing):1E)
mod:01(end(icl>time):29.@def,      century(icl>time):2H.@def)
aoj:01(21:2L,      century(icl>time):2H.@def){/unl}
[/S]
;;Done!
```

6. Conclusion

The UNL system maybe thought of as interlingua, but it has a number of other features that make it better suited for semantic inference than most of other interlinguas. In particular the following can be highlighted:

a) The set of Universal Words with universal interpretations ontologically built in information, e.g. cholera(icl>disease) which characterizes cholera as a type of disease.
b) It has a small and simple set of predicates with binary relations.
c) It includes a Knowledge Base connecting the Universal Words as a weighted graph of relations.
d) It is economic as the same dictionary is used in analysis and generation.
e) The dream of language independent semantic analysis.
f) The world wide efforts in developing mechanisms for converting language into UNL and vice versa, 15 language centers have been established so far.
g) UNL can replicate the functions of natural languages in human communications.
h) The UNL provides a mechanism for inferring.
i) UNL supports useful applications for multilingual society i.e. to represent contents written in any language and to generate any other language.

The EnCoding of Arabic structures in terms of hyper semantic networks in UNL format was completely feasible. It was possible to adapt Arabic morphology to UNL to extract concepts. Some challenges have been faced because of homographic ambiguities but they are controlled so far. Constructing semantic relations between the list of Universal words generated in the morphological stage was also possible. Constructing relations between modifiers first then between UWs representing the main skeleton of the sentence is proved to be a powerful design as it helped in controlling the application of the rules and prevented overlapping between them. Elaborated this way, the representation of Arabic structures in terms of UNL- based hyper semantic network could be accomplished which represents the encoding component in an interlingual system for man-machine communication in natural language could be built. This semantic network can be decoded back to any natural language given decoding rules and dictionary of the target language. Some problems are remaining e.g. selecting the correct possible segmentation of the word and dealing with collocations in the morphological phase, and automatic detection of constituent boundaries and controlling coordination in the relation phase which represent the main challenges of the future work.

7. References

Al-Ansary, S. (2003). Building a Computational Lexicon for Arabic, presented in the *17th* **ALS Annual Symposium on Arabic Linguistics.** 9-10 March 2003, Alexandria, Egypt.

Al-Ansary, S. and El-Kareh S. (2004a). Arabic - English Machine Translation Systems: Discrepancies and Implications. **JEP/TALN International Conference, Special session on Arabic text and speech language processing.** Fez, Morocco, 19-22 April 2004.

Al-Ansary, S. (2004b). A Morphological Analyzer and Generator for Arabic: Covering the Derivational Part. **NEMLAR International Conference on Arabic Language Resources and Tools**, Cairo, Egypt.

Alansary, S., Nagi, M. and Adly, N. (2006). Generating Arabic text: The Decoding Component of an Interlingual System for Man-Machine Communication in Natural Language, **the 6th International Conference on Language Engineering**, 6-7 December, Cairo, Egypt.

Arnold, D. (1994). **Machine Translation: An introductory Guide.** Manchester, NCC Blackwell.

Auh T. (2001). Language Divide and Knowledge Gap in Cyberspace: Beyond Digital Divide**. International Symposium on Language in Cyberspace**, 26 - 27 September 2001, Seoul, Korea.

Beesley, K. R. and L. Karttunen (2003). **Finite State Morphology**. Stanford, Calif., CSLI; [Bristol : University Presses Marketing, distributor].

Dorr, B. J. (1993). **Machine Translation: A View from the Lexicon**. Cambridge, Mass., MIT Press.

Eldakar Y., Adly N., and Nagi M. (2006). **A Framework for the Encoding of Multilayered Documents**, accepted for publication at First International Conference on Digital Information Management (ICDIM 2006), Bangalore, December 6-8, 2006.

Eldakar Y., El-Gazzar K., Adly N., and Nagi M.(2005). The Million Book Project at Bibliotheca Alexandrina, **Journal of Zhejiang University SCIENCE**, vol. 6A, no. 11, pp. 1327-1340, Nov. 2005. and in Proceedings of International Conference on Digital Libraries 2005 (ICUDL05), Hangzhou, China, pp. Nov. 2005. Available: www.zju.edu.cn/jzus/2005/A0511/A051122.pdf

Eynde, F. v. (1993). **Linguistic Issues in Machine Translation.** London, Pinter.

Fillmore, C. (1968). The Case for Case. In Bach, E. and Harms, R.T. (orgs.), **Universals in Linguistic Theory**, pp. 1-88. Rinehard and Winston, New York.

Galinski C. (2001). Dialogue among Civilizations in the Cyberspace**. International Symposium on Language in Cyberspace**, 26 - 27 September 2001, Seoul, Korea.

Hausser, R. R. (1999). **Foundations of Computational Linguistics: Man-machine Communication in Natural Language**. Berlin ; New York, Springer.

Hausser R. (2001). Human-Computer Communication in Natural Language. **International Symposium on Language in Cyberspace**, 26 - 27 September 2001, Seoul, Korea.

Hutchins, W. J. and H. L. Somers (1992). **An Introduction to Machine Translation.** London, Academic.

Johnson M. (1988). Where do I speak into it? A discussion of a method and motivations of Natural Language Processing. **Journal of Information Technology** (Routledge, Ltd), Vol. 3 Issue 3, p216.

Kiraz, G. A. (1996). **Computational Approach to Non-linear Morphology,** University of Cambridge.

Kiraz, G. A. (2001). **Computational Nonlinear Morphology : With Emphasis on Semitic Languages.** Cambridge, Cambridge University Press.

Klavans (1997). Computational Linguistics. In O' Grady W., Dobrovolsky M. and Katmba F. (eds), **Contemporary Linguistics: An Introduction.** Longman.

Koskenniemi, K. (1983). **Two-level Morphology : A General Computational Model for Word-form Recognition and Production**. Helsinki, University of Helsinki, Department of General Linguistics.

Montviloff V.(2001**).** Meeting the Challenges of Languae Diversity in the Information Soceity. **International Symposium on Language in Cyberspace**, 26 - 27 September 2001, Seoul, Korea.

Nirenburg, S., H. L. Somers, et al. (2003). **Readings in Machine Translation.** Cambridge, Mass. London, MIT.

Saleh I., Adly N. and Nagi M. (2005). DAR:A Digital Assets Repository for Library Collections, **9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)**, Vienna, pp. 116-127, Sep. 2005.

Uchida, H. (1996).**UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration**. UNU/IAS/UNL Center. Tokyo, Japan.

Uchida H. (2001). The Universal Networking Language Beyond Machine Translation**. International Symposium on Language in Cyberspace**, 26 - 27 September 2001, Seoul, Korea.

Uchida H., Zhu M. (2002a). Universal Word and UNL Knowledge Base**, International Conference on Universal Knowledge and Language (ICUKL)**, Goa, India.

Uchida H. (2002b). How to Build Universal Knowledge, **International Conference on Universal Knowledge and Language (ICUKL)**, Goa, India.

Uchida H.(2003). Knowledge Description Language, **Semantic Computing workshop,** Tokyo, Japan.

Uchida H., Zhu M. (2005). UNL2005 for Providing Knowledge Infrastructure, **Semantic Computing Workshop (SeC2005)**, Chiba, Japan.