International Journal of Computational Linguistics and Applications
vol. 7, no. 1, 2016, pp. 31–50
Received 27/06/2015, accepted 10/09/2015, final 23/09/2015
ISSN 0976-0962, http://ijcla.bahripublications.com

Realization of Minimum Discursive Units Segmentation of Arab Oral Utterances

CHAHIRA LHIOUI[1], ANIS ZOUAGHI[2], AND MOUNIR ZRIGUI[3,4]

[1] *Gabes University, Tunisia*
[2] *Sousse University, Tunisia*
[3] *Monastir University, Tunisia*
[4] *ESSTT Tunis, Tunisia*

ABSTRACT

Unlike the written texts, discourse segmentation of the Arab oral dialogues is a challenging task that is held back in most cases by the spontaneous character of oral speech. Like any segmentation task, segmentation in minimum discursive units (UDM) aims to cut the different statements of a speech into simple proposals easily usable in subsequent treatment. The majority of the work on the Arabic language was based on extensive syntactic analysis approaches. In this article, we try to show the effectiveness of hybrid approaches combining linguistic and probabilistic processes over purely linguistic approaches. The performance of our segmentation was evaluated on a relatively large size corpus. We built this corpus by using the method of the wizard of Oz.

KEYWORDS: segmentation, discursive unit, Arab oral statements, Wizard of Oz

1 Introduction

The first step in the analysis of a transcribed speech is its segmentation. It involves cutting, according to the analysis to be undertaken (seman-

tic, lexical, morphological or syntactic), statements in units of a certain type that will be previously defined in order to locate desired information. These units can be at different structural levels that we can quote respectively; the phrases, the proposals (what is called also the clauses or the minimum discursive units (UDM) [10], chunks, graphic words, lexical units (Word-shape [15]), morphemes, etc.

Segmentation has been described in several researches as a crucial stage prior to the linguistic treatment [9], because the quality of the final results depends on it. However, studies on segmentation are not numerous and not taken seriously by most laboratories which treat language automatically. In fact, each research team has developed an interim tool for well-defined corpus or has only used a manual processing. This lack is intensifying especially in Arabic where there is little work on the segmentation of written texts and there is virtually no functional and specific segmenter to the Arabic language in the context of an oral conversation.

In reality, the automatic segmentation of the Arab oral statements presents several difficulties linked on the one hand, to the specific characteristics of the Arabic language, and on the other hand, to the spontaneous nature of oral speech . Indeed, the specific morpho-lexical and syntactic characteristics of this language make it among the most difficult languages to control in the field of the NLP (Natural Language Processing). Its agglutinative nature, inflectional richness and the absence of vocalization generate a large number of virtual and actual ambiguities causing an important combinatorial explosion, especially at the level of the morphological analysis [15]. In addition, oral statements are in most of the cases uncertain and ambiguous. This adds difficulties to the automatic segmentation at the semantic level. In other words, UDM segmentation becomes increasingly intricate as the semantic coverage of oral statements is inadequate.

Given the intrinsic characteristics of the oral speech, we cannot be satisfied for simply designing a speech segmenter in an identical manner to a segmenter of a written text in standard Arabic language where in approaches based on punctuation marks and approaches by contextual exploration [14] [7] have no interest in the context of the oral.

2 Previous Works

The authors of [3] have proposed a rules-based approach for segmenting in sentences a non-vowel Arab text. The approach is a contextual analysis of the punctuation signs, conjunctions of coordination, and a list of particles, which are regarded as segmentation criteria. The authors identified 183 rules implemented by the STAr system. [12], in turn, suggested a rule approach guided only by lexical connectors (the punctuation is not taken into account) to segment the Arab texts in clauses. The authors introduce the concept of active connectors that indicates the start or the end of a segment and the concept of passive connectors that does not involve a point of break. The same connector can be passive or active in changing from one context to another.

The authors of [16] have proposed a learning method for the segmentation of the Arabic texts in clauses using only the rhetorical functions of the connector "و/et. The authors have defined six senses for this connector such as: Al القسم (Aloquasam), الإستئناف (Alo < isti'onAf), العطف (AloEaTof), etc.

The authors of [13] have used a rule-based approach for the segmentation of the Arabic texts in clauses. In this work, three principles of segmentation have been used. The first principle uses only the punctuation signs. The second one relies on lexical indices. The third principle combines the punctuation signs and lexical indices in order to address the ambiguities of lexical indices.

The authors of [10] has shown the feasibility of UDM discourse segmentation for modern standard Arabic language (ASM) under the SDRT theory. To do this, they used a supervised learning multi-class method that predicts the embedded UDM. To our knowledge, it was the first work on the segmentation of the Arabic texts into discursive units.

3 Segmentation Types

There are several levels of analysis that we can focus on to identify the various elements constituting the text and define the borders. We can stop at the sentence level or the proposal or the phrase. But we can also achieve the level of graphic Word, lexical units or go beyond them to arrive at the base units that they compose which are morphemes. In fact, according to the intent of the analysis to be undertaken: lexical,

morphological or syntactic, we can classify the segmentation in three application types:

- The itemization (tokenization or Word segmentation) which is the segmentation of text into words or lexical items (tokens). This type of segmentation is also called lexical segmentation. It comprises morpho-syntactical treatment (labeling or POS tagging in English).
- Morphological segmentation that goes further than the lexical segmentation seeking to isolate the different components of lexical items in separate smaller units, which are the morphemes.
- The chunking is to isolate the different components of the text into independent unit longer than words and less than the proposed units. These are called phrases. This type of segmentation is also called syntactic segmentation.
- Discourse segmentation means segmentation in simple proposions written texts. A simple proposition is a full sentence that has a definite meaning. The propositions may be affixed or contingent proposals or also correlated phrases etc. [10].

4 Difficulties of the Arab Oral Statements Segmentation

The automatic segmentation of the Arab statements presents several difficulties related to the specific characteristics of the Arabic language as well as to the inherent characteristics of the oral.

4.1 Common Properties Between the Written Arabic and the Spoken Arabic Inhibiting Discourse Segmentation

Agglutination One of the difficulties caused by discourse segmentation is the phenomenon of agglutination. Indeed, the words can have an agglutinated structure resulting from a concatenation of lexical and grammatical morphemes. Thus, a word in Arabic may represent a proposition. The following example shows the structure of a verb in Arabic agglutinated form:

أستزورونا

Are you going to visit us?

Generally, the agglutinative structure of a word is formed by a sequential concatenation of three components which are read right to left, respectively:

Enclitic	suffix	root	prefix	proclitic

Fig. 1. The diagram of an agglutinated structure of a verb in Arabic language

Firstly, there are four prefixes gathered in the word 'أنيت' (cf. Table 1.), represented by a single morpheme each corresponding to a single letter at the beginning of a word. They indicate the person of the verb conjugation in the present tense. We can only use one single prefix at the same time. However, these prefixes may coincide with the letters forming the word. For example, the first letter of the word أكل coincides with the prefix أ which indicates that the verb is conjugated in the first singular person present.

Table 1. Arabic prefixes and their designation

أ	Means the first singular person (I)
ن	Means the first person in plural (we)
ي	Means the third masculine singular person, dual, plural, masculine and feminine plural هو(He, It), هي (She), (They) هن, هم
ت	Is the second person feminine, masculine, singular and dual

Suffixes are also considered as another source of ambiguity similar to prefixes. In fact, they have a double interest:

–Termination of the conjugations of verbs

–Designation of the gender of plural and feminine names

Due to the dual role, suffixes can be confused with the letters of the word itself. For example, the letter و in the word يحبو (he crawls) does not refer to the plural form of the verb يحب (he loves).

Finally, proclitics combined together give more information about the Arabic word (semantic traits, coordination, determination...). Here are some examples of proclitics:

The coordination by the conjunctions: ف 'fa' and و 'wa '.

–The brand of the future: س ‘sa’

–Article: ال 'al'

–Prepositions with the letters: ب 'bi' and ل 'li'

– Querying: أ

However, in writing, it is not always easy to tell the difference between a proclitic and a character belonging to the root of certain words. For example the character س in the word سبح (he bathed) is a character from the verb root; in the word سأقضي (I'll spend) it is a proclitic that marks the future. In addition, there may be confusion between a brand of coordination (و) and a character from the root وجد (he found) or a suffix designating a sign of plural حجزوا (they have booked).

Infrequent Use of Vowels There is another difficulty in the Arabic writing system which is the optional use of vowels that are added above or under the letters in the form of diacritics (see Table. 2). The disjunction between consonants and vowels is a source of ambiguity which hinders the discourse segmentation of the sentences. These signs are useful for the understanding of the sentences and thus help dissecting the simpler proposals independently of the position of the words in the sentence. The proportion of ambiguous words attains over than 90% if the counts relate to global voyellation (lexical and casual) of these words [6].

Table 2. Diacritics

الفتحة ◌َ	الكسرة	الضمة ◌ُ
فَ	فِ	فُ

The following table provides an example for the word حجز (hjz under the transliteration of Backwalter [4]) writing in the non-vowel form.

Table 3. Variation of the diacritic signs of the word حجز -hjz and its different interpretations

Interpretation (3)	Interpretation 2	Interpretation 1	Word without vowel
حَجْزٌ hajzun « booking»	حُجِزَ hujiza "it has been booked"	حَجَزَ hajaza "he booked"	حجز hjz To book

Inflectional and Derivational Ambiguity One of the specificities that characterize the words of the Arabic language is that they derive from a root to three radical consonants in the form of (فعل) [2]. These derivations are obtained by using the combination of different schemes. Arabic offers a total of 150 schemes some of them are complex as the repetition of a consonant (with the diacritic ّ الشدة (Achaddah)) or the elongation (we talk about the long vowels (see Table 4.) of the root consonant. This phenomenon is a characteristic of Arabic morphology which makes this language morphologically, syntactically and semantically rich and ambiguous and also resulting in huge difficulties at the time of the automatic segmentation into sentences (UDM). Indeed, the identification of the grammatical category of words becomes ambiguous. This ambiguity may be intensified when it comes from a non-vowel word. For example, for the root "حجز" (to book), we have a derived form "حَجَزَ" which can designate a verb (he booked) or a noun " حَاجِزٌ" (person who book)).

Table 4. Long vowels

ـِي المدّ بالكسر	ـُو المدّ بالضم	ـَا المدّ بالفتح
فعِيل	مفعُول	فَاعل

Structural Ambiguity The Arabic sentence can be either simple containing a single proposal, or complex admitting more than a juxtaposed proposition, coordinated or connected by conjunctions of subordination.

A proposition can be either verbal, starting with a verb or nominal starting with a name. The verbal proposition contains a single verb and one or several subjects at which it may be added one or more object complements when the verb is transitive (example: 'أسافر' (I travel): (intransitive verb). While in the nominal proposition, we talk about the theme (المبتدأ) and the proposal (الخبر).

Generally, phrases in Arabic are long and can reach a whole page in some cases. This can engenders high complexities in segmentation and understanding.

Lack of Capitalization and Punctuation Unlike Indo-European languages, Arabic language doesn’t use capitalization, which complicates

the determination of the segment boundaries. In addition, the punctuation is not used in a systematic way which is the case for French and English. This greatly complicates the task of discursive segmentation [10].

4.2 Specificities of the Oral

One of the difficulties sources, when using the UDM transcribed speech segmentation, is linked to the particularity of the oral modality and the spontaneous nature of the interaction.

Oral Modality The syntax of the oral is different and less strict than that of the written. The statement may make some ungrammaticality. The following example shows that the second replica of the user does not undergo the usual form that an Arabic phrase must have (Verb-Subject-Complement or Theme-Proposal):

(User) - أريد السفر إلى سوسة

(System) - نعم، وماذا أيضا

(User) - وإلى قربص

This paralyzes a discourse segmentation of this statement.

Oral Spontaneity When the user speaks spontaneously, the statement may necessarily includes the hesitations, slips of the tongue, false departures, corrections or repetitions of words or some groups of words and other phenomena which disturb the segmentation task. The user can also interrupt the system, which causes a problem of speech recovery. For example, in the sentence "أريد معرفة توقيت القطاريل ثمن تذكرة القطار" (I want to know the train table time but rather the ticket train price), the word sequence "ثمن تذكرة القطار" (ticket train price) is related to the verb "أريد" (I want) which is located in the segment"أريد معرفة توقيت القطار" (I want to know the train table time). The latter have normally to be rejected during the pretreatment phase since it is a false start.

Ellipsis Phenomenon One of the most important topics which complicate the task of segmenting is the phenomenon of ellipse [11]. This phenomenon appears frequently as well in the oral as in the written

texts. It consists in the omission of words, some expressions and even a sentence. Below is an example illustrating an ellipse of an expression in the sentence:

عاما سعيدا [أتمنى لك]: [I wish] (good year): an ellipsis of *wish* expression

Non-fixed Word Order Most frequent difficulty in oral written inhibiting UDM segmentation is the non-fixed order in statements. Indeed, the change in position of certain words does not necessarily change the meaning of the sentence. For example, in the case of a nominal phrase, if the proposal starts with a preposition, the theme/proposal positions can be reversed (هناك سمكة في البحر (There is a fish in the sea) becomes ' هناك في البحر سمكة ' (There is in the sea a fish)).

Wrong Transcription Finally, sentence segmentation is greatly influenced by wrong transcription errors which reached to 66 percent error rate according to [8]. Indeed, owing to the rigid specificities of some Arabic letters such as: strong expiration letters (ع،ح، خ، ق،غ etc.) respectively (ghayn, Ha, qaf, Kha, ayn) where an ambiguity in the segmentation can be obtained. That's to say, the word "أتعلم" (did you know) for example can be transcribed by the word "أتحلم" (dreaming) because of the confusion between the two letters 'ح' and 'ع' since these two letters are close in their pronunciation (two glottal sounds).

5 Our Method

According to the existing study, we have noticed that there does not exist any specific work, which treats discourse segmentation of the Arabic oral statements. In fact, all works have focused on written texts. To segment a text written in Modern Standard Arabic (MSA) is easier than segmenting a transcribed MSA oral statement. Firstly, the nature of the written Arabic is intrinsic. Secondly, due to the spontaneity of oral communication, oral language is ungrammatical and non-deterministic. Giving that an UDM constitutes a grammatically and semantically complete proposal, an UDM of an oral transcription is more delicate.

In conclusion, added to the difficulty posed while applying the discourse segmentation in the standard language in the case of writing, another difficulty arises in oral segmentation. Hence a similar task would also be difficult. In this context, we fit our work in order to find an effective discourse segmentation strategy dealing with MSA oral statements.

5.1 UDM Segmentation Process

To segment an MSA transcribed oral statement, we opted for an approach in three basic steps:

- First, we perform a morpho-lexical segmentation. This segmentation is done in two separate steps. The first performs a coarse lexical segmentation (tokenization) based only on spaces. With regard to the second, it exerts a more detailed morphological segmentation dealing with the case of agglutinated words.
- Second, we remove intruder obtained words (such as duplications and unnecessary information (like: 'آم' (Eum), 'آه' (Euh) interjections). We also keep small interests to the elliptical forms and fault departures. Then, we convert numbers written in all letters, and to determine the canonical forms of words. This second step can be summarized on a pretreatment word.
- During the latter step, we examine the obtained pretreated tokens (also called lexical units), and we compare them to the already segmented and pretreated forms of a semi-automatically pretreated corpus:
  - If the token is found in the corpus, its segmentation is validated.
  - If it is not, we search by using a regular expression that represents the complete form of an Arabic word (pre-bases/root/post-bases), the possible pre-bases and post-bases attached to the root, where pre-bases and post-bases designate respectively the couples (proclitic, prefix) and (enclitic, suffix). This regular expression is built from lists defined in advance. For each identified pre-base and post-base, we check the status of the remaining part of the cut-out word.
- Next, we move to a syntactic labeling step of lexical units extracted during the first step.

– Finally, we regroup tokens obtained to form UDM that represent the user intensions. This is done by using Probabilistic Context-Free Grammar (PCFG) rules established from the corpus.
–

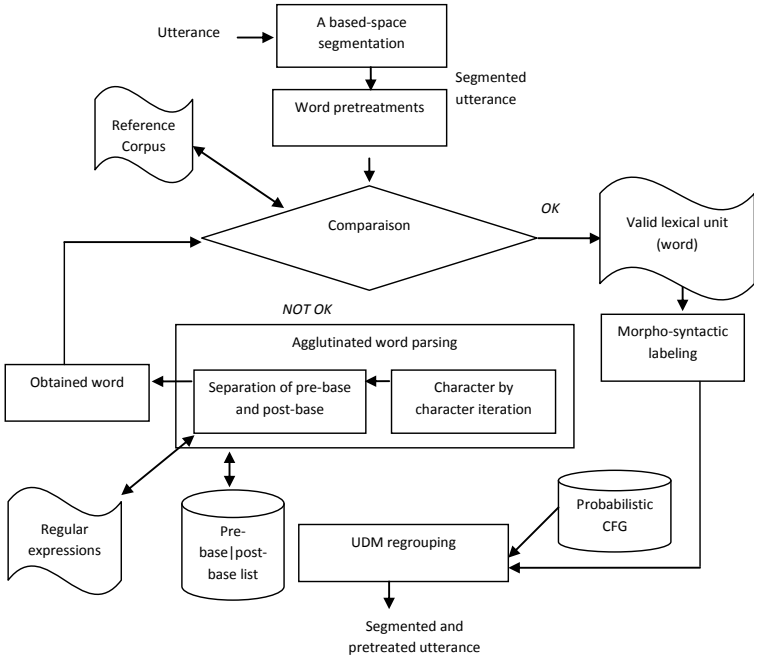


Fig. 2. UDM segmentation process

As it is shown in Fig 2., we have tried to develop a discourse segmentation algorithm based on a hybrid approach involving linguistic and probabilistic techniques by using the PCFG grammar.

The success of our method is essentially based on a sufficient list of regular expression and language rules as well as on a large list of annotated and segmented dialog words. This list forms what we call reference corpus.

A regular expression is a pattern that describes the complete form of a word:

Pre-base	Root	Post-base

where pre-bases define (prefix + proclitic) and post-bases represent (post-fixes + enclitic).

Concerning linguistic rules, they are automatically extracted from the training corpus (see section 5.2.).

5.2 Stochastic Grammar Learning

We have chosen the context free grammar type because it is less restrictive and more adaptable to the irregularities of the oral than other types of restrictive formal grammars. Two problems could be detected during the grammar learning:

- Learning of grammar rules,
- Learning of associated rule probabilities.

Our grammar rules are learned from a training corpus. The latter have to be represented in a tree form. Indeed, the task of learning grammar rules is considerably facilitated when linguists (or even native speakers trained) analyze the data in syntactic trees. The creation of such trees corpus represents a big investment. Thus, in our case, we are based on our own corpus of tree established and described in our work reference by [5]. We create a PCFG by counting: for each non-terminal symbol, we just look at all the nodes that have this symbol for root and create rules for each different combination of leaves in these nodes. For example, if GVerb symbol appears100 000 times and if there are 20 000 GVerb instances with the list of nodes [GAdv, GVerb] we create the rule:

$$\text{GVerb} \rightarrow \text{GAdv GVerb}$$

We have chosen to model our grammar by a Hidden Markov Model (HMM) due to the clarity, rigor, efficiency and generality that it presents. Hidden states from our model represent syntactic trees of the training corpus. Therefore, the problem of learning the established grammar probabilities is reduced to a basis of HMM-supervised learning. That's to say, it is a training problem that consists in estimating digital parameters (first visit probability distributions, transition and generation) in a way to explain well the learning sequences. We have adopted an EM (Expectation-Maximization) approach to estimate probabilities. Step E estimates the probability of each sub-sequence that is

generated by each rule, and then the M step estimates the probability of each rule. All the calculation can be done by dynamic programming using an algorithm called the Backward-Forward (BF) algorithm of HMM models.

We used a second method for estimating probabilities which is the method of maximum likelihood (Maximum Likelihood: ML). Having a PCFG grammar, and from the positive examples (properly constructed sentences) which constitute the corpus in a tree form called training corpus, we can easily estimate the probability of each rule in the grammar by the ML method. The form of such a probability is noted as follow:

$$q_{ML}(X \rightarrow w) = \frac{count(X \rightarrow w)}{count(X)}$$

with $(X \rightarrow w)$ denotes the number of times where the rule $X \rightarrow w$ is encountered in the corpus and $count(X)$ represents the number of times where the non-terminal X is encountered in this same corpus. For example, if the rule $VP \rightarrow Vt\ NP$ is cited 100 times, in the corpus, and non-terminal VP is met 1000 times, then

$$q_{ML}(VP \rightarrow Vt\ NP) = \frac{100}{1000} = 0.1$$

Ambiguities Treatment We noticed the presence of the ambiguities in the agglutinated case. These can be segmented in several different ways. This is due to the ambiguous nature of the Arabic language including the use of particles as the 'و' (*wa*) which means sometimes a coordinating conjunction, a part of the word or a sign of plural. If it (the 'و') is a plural it designs a letter and it is putted at the end of the word. Otherwise, if it is a conjunction word it is at the start. It would be attached to the previous or following word, respectively, except for non-connecting letters such as the 'ف' (*fa*).

Morphological ambiguities create ambiguities of the higher level (lexical semantics and even pragmatic). For example the Arabic word المهم can be segmented in five different ways depending on its context in the sentence (see following table):

Table 5. Different divisions and interpretations of the word 'المهم'

Possible cutting	Translation into English
المُهِمُّ	interesting
الَمّ + هَمٌّ	sake + pain
أَلَمْ+هُمْ	they + is what
أَلَمٌ +هُمْ	their + pain
أ+لِمَ+هُمْ	they + it + is

This problem remains difficult to solve since these types of words segmentation depends necessarily on context and its position in the sentence. In this case, our segmentation algorithm first takes all of the agglutinated word and cut it using a regular expression. Then it compared to existing words in the corpus holding the valid cutting. That's why the quality of the segmentation depends on the size of the corpus that is supposed to contain the most frequent words in Arabic with their correct segmentation.

5.3 Reference Corpus

We built and used our own corpus. This corpus is dedicated to the study of the applications of demand for hotel reservations, tourist information. The dialogues are designed to the booking of one or more rooms in one or several hotels. Bookings are made within the organization for a weekend, holiday or a business stay.

Corpus Collection These dialogues were collected using the Protocol of the wizard of Oz (Wizard of Oz, WoZ) [5]. During the interaction, users believe conversing with a machine while the dialog is actually supported by a human operator that simulates responses from a server information and booking. The operator is assisted by the WoZ tool in the generation of responses to provide to the user.

After each user phrase, the operator refers to the WoZ tool which offers the answer to provide on the basis of the new state of the dialog. To diversify the operator answers, the WoZ tool is set to the level of messages, instructions and scripts. A set of messages is associated with the application to vary the formulations of answers. At each call, the operator must comply with a series of instructions (for example, pre-

tending not to have understood the user to simulate the errors that would make a real system). These instructions must be provided to the WoZ tool and depend on the scenario chosen for the dialogue to save. The table below shows general characteristics that qualify our corpus.

Table 6. Statistics from our corpus

Complexity indices	Value
Utterances number	1000
System users	100
Queries type	14

Segmentation and Corpus Annotation To properly ensure the supervised learning based on HMM of linguistic rules in our system, we have divided our corpus in two illegal parties: usually, 2/3 of the corpus is reserved for the training corpus (TRN) and 1/3 for the test (TES). We have established a phase of segmentation and manual annotation by two Arab native annotators. These have annotated the corpus of training according to the guidelines set out in the manual of annotation of AlKhalil [1]. We get a kappa of the inter-annotator agreement of 0.87% for our training corpus.

5.4 Regrouping on UDM

Once cut and labeled, the words will be grouped in order to restore the minimum discursive units covering them. Given the HMM of the application, the problem of the UDM construction is reduced to the problem of decoding. The later consists on the determination in optimal way of the hidden component Q of the stochastic process, given the observable component O, and probabilistic information about HMM model noted H. What concerns us here is not the value of the maximum likelihood but the path or a series of states that maximizes $P_H(Q|O)$ that is called the criterion of maximizing. According the Bayes rule, maximize $P_H(Q|O)$ is to maximize the amount of $P_H(Q,O)$. The solution of this criterion is called Viterbi states sequence because it is found using the Viterbi algorithm.

6 Evaluation and Results

Taking into consideration of the good agreement results, the three annotators were invited to build our corpus by consensus. A total of 140 dialogues for the training corpus, we have a total of 70 UDM in which 10% UDM are embedded and 30% UDM are scattered on more over than one statement, 20% of agglutinated words. Similarly, for simplification reasons, we have kept virtually the same percentages for the test corpora.

Table 7. Reference corpus characteristics

	Word/Dialogue	UDM	Imbedded UDM	Dispersed UDM	Agglutinated words
Training corpus	9800/140	70	7	21	1960
Test corpus	3500/70	30	3	9	700
Total	**13300/210**	**100**	**10**	**30**	**2660**

The results obtained in the table below show that measures calculated for a stochastic grammar are better than those calculated for a non-stochastic grammar while we omit rule probabilities. In addition, the measures that were found for probabilities estimated by the method of ML are better than those calculated for a grammar estimated by the method of Backward-Forward. Thus, ML method brings a good discourse segmentation which is useful in further processing.

According to the table in Table 8., the CFG reveals a weakness in the detection of the scattered or dispersed UDM. This can be explained by the ambiguous cases that the probability distribution can deal with. Besides, oral intrinsic criteria including its spontaneity and the extensive existence of incomplete and unfinished sentences make the non-probabilistic CFG more incapable to succeed segmentation phase.

Table 8. Detection percentage of dispersed and scattered UDM and agglutinated word by the linguistic and stochastic approach

		Dispersed UDM		Embedded UDM		Agglutinated words	
	Corpus	**ENT**	**TES**	**ENT**	**TES**	**ENT**	**TES**
CFG		0.2	0.01	0.367	0.289	0.196	0.2
PCFG	Estimate by Back-word-Forward (BF)	**0.79**	**0.615**	0.231	0.346	**0.431**	**0.402**
	Method of likelihood (ML)	0.406	0.389	**0.598**	**0.699**	0.197	0.201

Indeed, the incomplete and unfinished sentences are principally those which form the dispersed UDM. The following examples illustrate scenarios of dispersed UDM on two replicas caused by a non-achieved idea in the first utterance.

(UDM1) : [أريد حجز نزل ليلة 17 من ديسمبر [

[I want to book a hotel on the night of 17 December [

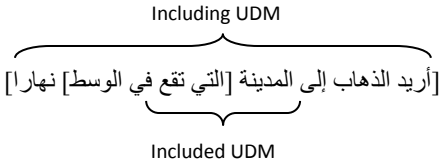
وماذا أيضا؟

And what next?

(UDM2): [18 و 19 ديسمبر]

[18 and 19 December]

The second example shows how the UDM nesting are.



Moreover, the ambiguities caused by some particles through the adding of prefixes and suffixes (see example below) disrupt the UDM detection by the PCFG estimated via the method of BF. This case is generally resolved by the analysis of the discourse context.

Example:

– Temporal particles: قَبل (in front of), can be linked with the prefix 'ف' (then) to become فقبِل which causes an ambiguity specifically when there are no signs of diacritics. Thus, instead of being interpreted as a separator of UDM segment, it will be confused with the verb *come*.

7 Conclusion and perspectives

In this paper, we presented a discourse UDM segmentation of the Arab oral dialogues task. This phase is held back in most cases by the spontaneous character of oral speech while dealing with an Interactive Voice System. Like any segmentation task, the UDM segmentation aimed to cut the different statements of a speech into simple propositions easily usable in subsequent treatment. We also showed the effectiveness of hybrid approach combining linguistic and probabilistic processes using the probabilistic grammar (PCFG). We compared PCFG results with purely linguistic approach using CFG and we found that the second approach was less effective than the first one. The performance of our segmentation was evaluated on a relatively large size corpus built using the wizard of Oz method.

As a perspective, we expect to ameliorate our segmentation component to take more consideration to the UDM dispersed cases.

References

1. Alkhalil Morpho Sys, Version 1.0, 2010, http://www.alecso.org.tn/index.php?option=com_content&Itemid=956&lang=at
2. Baloul.S,Alissali.M, Baudry.M and Boula de Mareuil.P. Interface syntaxe-prosodie dans un système de synthèse de la parole à partir d'un texte en arabe. *24es Journnées d'étude sur la parole*, 329-332, 2002.
3. Belguith H. L., Baccour L., et Mourad G., Segmentation de textes arabes basée sur l'analyse contextuelle des signess de ponctuations et de certains particules, 12th conference on Natural language Processing, TALN'2005, Dourdan.
4. Buckwalter.Tim, Arabic Morphological Analyser version 1.0 Linguistic Data Consortium. Catalogue numéro LDC 2002 L49, 2002.
5. Chahira L., Anis Z., Mounir Z., A combined A Combined Method Based on Stochastic and Linguistic Paradigm for the Understanding of Arabic

Spontaneous Utterances, A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 549–558, 2013.

6. Debili F., Achour H., Souissi E. : « La langue arabe et l'ordinateur, de l'étiquetage grammatical à la voyellation automatique », *Correspondances n° 71* juillet-août 2002, Site Internet : www.irmcmaghreb.org
7. Descles J.-P., Systèmes d'exploration contextuelle. Co-texte et calcul du sens., éd. Claude Guimier, Presses Universitaires de Caen, pp. 215-232, 1997.
8. Gordon, Joshua, Rebecca J. Passonneau, and Susan L. Epstein. "Learning to balance grounding rationales for dialogue systems." Proceedings of the SIGDIAL 2011 Conference. Association for Computational Linguistics, 2011.
9. Habash.N and Rambow.O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambguation in One Fell Swoop. In Proceedings of the Conference of the Association for Computational Linguistics Ann Arbor, MI, 2005.
10. Iskander K., Farah B., Lamia H., « Segmentation de textes arabes en unités discursives minimales », *TAL-RECITAL,* Les Sables d'Olonne du 17-21 juin 2013
11. Kais H. Abdelmajid B., An Ellipsis Resolution System for the Arabic Language, International journal of Computer Processing of Languages, volume 22 number 4, décembre 2009
12. Khalifa I., Feki Z., Farawila A., Arabic Discourses Segmentation based on Rethorical Methods, International Journal of Electric and Computer Sciences, IJECS-IJENS, Vol: 11 (1) , 2011.
13. Keskes I., Benamara F., Belguith L., Clause-based discourse Segmentation of Arabic Texts, the eighth international conference on language resources and Evaluation, LREC Istanbul, 21-27 may 2012.
14. Leïla Baccour, Lamia Belguith Hadrich , Ghassan Mourad, Présentation d'un Segmenteur de Textes Arabes : STAr, 2004.
15. Mouelhi Z., AraSeg* : un segmenteur semi-automatique des textes arabes, JADT'08 : 9es Journées internationales d'Analyse statistique des Données Textuelles, 2008.
16. Touir A, Mathkour H., AL-Sanea W., Semantic based segmentation of Arabic texts, Information Technology Journal Vol: 7 (7), 2008.

CHAHIRA LHIOUI
ISIM OF MEDENINE,
GABES UNIVERSITY,
ROAD DJERBA, 4100 MEDENINE, TUNISIA
E-MAIL: <CHAHIRA_M1983@YAHOO.FR>

ANIS ZOUAGHI
ISSAT OF SOUSSE,
SOUSSE UNIVERSITY,
TAFFALA CITY (IBN KHALDOUN), 4003 SOUSSE, TUNISIA
E-MAIL: <ANIS.ZOUAGHI@GMAIL.COM>

MOUNIR ZRIGUI
FSM OF MONASTIR,
MONASTIR UNIVERSITY,
AVENUE OF THE ENVIRONNEMENT 5019 MONASTIR, TUNISIA
AND LATICE LABORATORY,
ESSTT TUNIS, TUNISIA
E-MAIL: <MOUNIR.ZRIGUI@FSM.RNU.TN>