Spoken Egyptian Arabic Word Segmentation using Phonotactics

Reham M. Marzouk, Alexandria University
marzoukreham@gmail.com

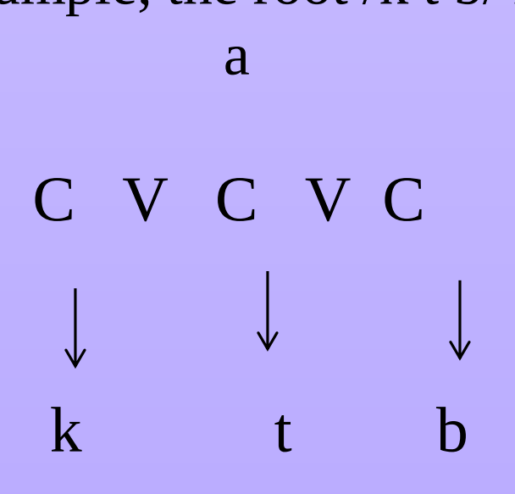
1. INTRODUCTION

This study is a modeling of the Egyptian Arabic phonotactics, it is based on the gradient constraints of the verbal root consonants sequencing. OCP-place (the obligatory contour principle of place of articulation) approach gave a large description of the Arabic triliteral root ,and the similarity constraints of its consonants, Since the Egyptian Arabic (EA) Dialect has been considered as the most widely understood dialect throughout the Arabic world, creating a model for spoken word segmentation of EA became an essential task.
The main goals are i) to formalize a model for the human phonotactic constrains detecting based on probabilistic measurements of the verbal root consonants, ii) to improve a segmentation performance according to these probabilistic measurements, and iii) to evaluate the ability of the model to detect the morphological tokens of a speech by evaluating the phonotactic constraints. the goals have been achieved by designing a phonotactic model measure the maximum likelihood for the verbal root consonants occurance based on the Hidden Markov Model states, and use the estimated values to detect the tokens boundaries.
Although the model is sill in progress its performance produced a high accuracy for the required assignments, and gave a high coverage for the morphological forms based on the phonotactic constraints.

2. ARABIC NON-CONCATENATIVE STRUCTURE

Arabic language is a member of the Semetic languages family which is distinguished by its root-pattern system, where verbal root consists of a set of three or four consonants.
Three approaches describe the Arabic word structure:

In **Templatic approach** (McCarthy, 1990) the Arabic word is represented in separate levels which are called Autosegmental tiers:
The root tier consists the consonants that construct the root
The pattern (skeletal tiers) represents the word in term of consonants C and the vowels V which are inserted together in a certain patterns or templates.
The vocalism (vocalic melody tier) represents the short vowels that specify each template or pattern.
For example, the root /k t b/ has among its word forms /katab/ 'to write' , /kutib/ 'to be written.and /kitaab/ 'book'



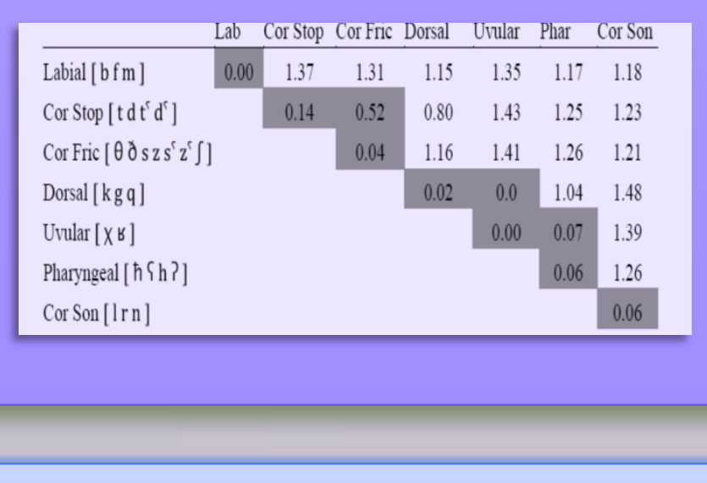
Affixational approach:
McCarthy and Prince (1990) claimed that templates is not the main issue in analyzing Arabic but measures are derived by affixation under what they had called the "prosodic circumscription".
Prosodic circumscription defines the domain of the morphological operation this domain is the grammatical category resulting by affixation.
Description of Egyptian Arabic Phonology:
EGY as MSA has a phonological profile includes 28 consonants, three short vowels, and 3 long vowels in addition to two diphthongs /aw/, /ay/
Egyptian Arabic followed the same phonological system of MSA but differs in the following:
MSA velar /q/ is realized in EGY as glottal stop />/
MSA alveolar affricate /d/ is realized in EGY as velar /g/
MSA coronal Obstruents /ð/, and /ɵ/ are realized in EGY as /d/ or /z/, and /t/ or /s/, respectively :
MSA diphthongs /aw/ and /ay/ have mostly become /o:/ and /e:/, respectively
Syllabification
The phonological word in the EA consists of a stem and affixes, such as definite articles /il/, conjunctions /wi/ subject and object pronouns
When morphemes cannot be syllabified according to the syllabificastion algorithm one of the following repairing processes occurs
Epenthesis:
When 3 consonants are juxtaposed within the utterance, an epenthesis of [i] or [u] occur between the second and the third
kull+naa……kull[i]naa
Consonant prosthesis
Although each syllable in CA requires an Onset, and the majority of morphemes which may occur in utterance-initial position have an underlying initial consonant (Watson, 2007), there are some vowel-initial morphemes such as the definite article /il/, at this case an onset is added through prosthesis of a glottal stop which is replaced in the continuous speech by the coda of the previous syllable.
Closed syllable shortening
When A domain-final syllable CVVC is concatenated with a consonant-initial morpheme, the long vowel is shortened (Watson,2007

3. PHONOTACTIC MODELING OF ARABIC LANGUAGE

Arabic root is a perfect example of OCP (Obligatory Contour Principle for Place of Articulation), while Pierrhamburt (1997), found that there is a relation between the consonant pairs within the root and the acceptability of the verbal root, this relation gradiently dependents on the homorganic similarity of these consonants, a highly similar homorganic consonants are less frequent than dissimilar homorganic consonant pairs
OCP-Place constraints are filters which prohibits a root with repeated homorganic consonants, but these constraints are non-categorical but a gradient constraints which have fuzzy constraint boundaries.
OCP place model (pierrhamburt 1997) is a stochastic constrain model that can be parameterized to account for gradient constraints within successive pairs of the Arabic verbal roots using the O/E ratio
O/E Ratio is the ration of the observed consonant pairs to the number of consonants expected to occur by chance (Pierrehumbert, 1993):

$$\frac{O(xy)}{E(xy)} = \frac{Prob(xy)}{\Sigma Prob(xY) \cdot \Sigma Prob(Xy)}$$

Where Y can be any segment following x, and X can be any segment preceding y.
The ratio of less than 1 indicates underrepresentation of the dataset as seen in fig(1)



	Lab	Cor Stop	Cor Fric	Dorsal	Uvular	Phar	Cor Son
Labial [b f m]	0.00	1.37	1.31	1.15	1.35	1.17	1.18
Cor Stop [t d tˤ dˤ]		0.14	0.52	0.80	1.43	1.25	1.23
Cor Fric [θ ð s z sˤ zˤ ʃ]			0.04	1.16	1.41	1.26	1.21
Dorsal [k g q]				0.02	0.0	1.04	1.48
Uvular [χ ʁ]					0.00	0.07	1.39
Pharyngeal [ħ ʕ h ʔ]						0.06	1.26
Cor Son [l r n]							0.06

4. METHODOLOGY

Phonotactic modeling
Corpus: Triliteral roots of 915 Egyptian words have been extracted from ARZ-ATB Arabic Treebank corpus, the corpus presents Egyptian written conversations
Transcription: CODA(Conventional Orthography for Dialectal) is used for transcription, CODA describes the Egyptian Arabic phoneme, it doesn't concern with the phonemic variations between speakers.
Hidden Markov Model
The Hidden Markov Model(HMM) is a powerful statistical tool for modeling generative sequences that can be characterized by an underlying process generating an observable sequence(Blunsom,2004)
Maximum Likelihood Estimate MLE was calculated for the most probable hidden state sequence of the adjacent root consonants C1C2, C2C3
.by the matrix:

$$\boldsymbol{p(c_n|c_{n-1}) = \frac{Count(c_{n-1,}\, c_n)}{Count(c_{n-1})}}$$

Where **c** is the consonant
Syllabification
First stage is to divide the transcribed text into syllables which follow the syllabic structure of EA
CV/CVC/CVV/CVVC./CVCC
The output is the target data that is used to detect the word boundary
Normalization
this stage aims to omit the repairing processes, like consonant prosthesis, i.e.morphemes , that use the final consonant of the previous syllable as an onset, are detected and the final consonant are replaced to its original position
Ex. mak.ta.buh (his office)……mak.tab.uh
Word boundary detecting
Aiming to detect the word boundary by determining the morphemes and the final syllable of the stem, MLE is used to estimate the probable sequence of root consonants which are proceeded by a boundary
Evaluation
A transcribed test data of 178 root-based words, (spoken and written continuously), were evaluated to test the Accuracy of the model and its ability to distinguish the stem and the morphemes boundaries
The data was divided into three sets
Data set1
Derived words, morpheme free
Data set2
Derived words, suffixed by one morpheme
Data set 3
Derived words prefixed by one morpheme
Testing metrics
I used the elements of ***confusion metrics*** which counts and visualizes when the system is confusing two classes and it computes the accuracy, precision, recall, and F score of any system, these elements are (true positive, false positive, true negative, and false negative)

Precision: Fraction of the item retrieved by the system that are interesting to the user

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall: Fraction of the items of interest to the user that are retrieved by the system

$$\text{Recall} = \frac{tp}{tp + fn}$$

Confusion metrics	Relevant	Irrelevant
Retrieved	True positive TP	False positive FP
Not retrieved	False negative FN	True negative TN

F score It is the mean of recall and precision, and it is interpreted as the weighted average of the recall and precision

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Accuracy: The accuracy is defined by the percentage of the correct predictions of a certain morphological feature category

$$\text{Accuarcy} = \frac{\text{TP+TN}}{\text{Total Number of morphemes}}$$

5. RESULTS

Three different thresholds have been evaluated and the best results have been Shown in the threshold 0.03%
The model performed good percentages of recall,, precision and accuracy, table (1)
The phonotactics of the root consonant succeeded to determine the boundary of the word stem from its suffixes and prefixes with an acceptable percentage.
The highest accuracy was for the data set three due to the observed Ability to detect the prefixes.
The model covered 82.22% of the data set prefixes and 44.74 % of the suffixes. fig (3)

Sample	Precision	Recall	F- score	Accuracy
Data 1	68.81%	82.0%	66.84%	67.0%
Data2	78.94%	81.81%	80.34%	72.5%
Data3	90.96%	90.96%	90.96%	85.84%
AVG.	79.57%	84.92%	79.57%	75.11%

Table (1)

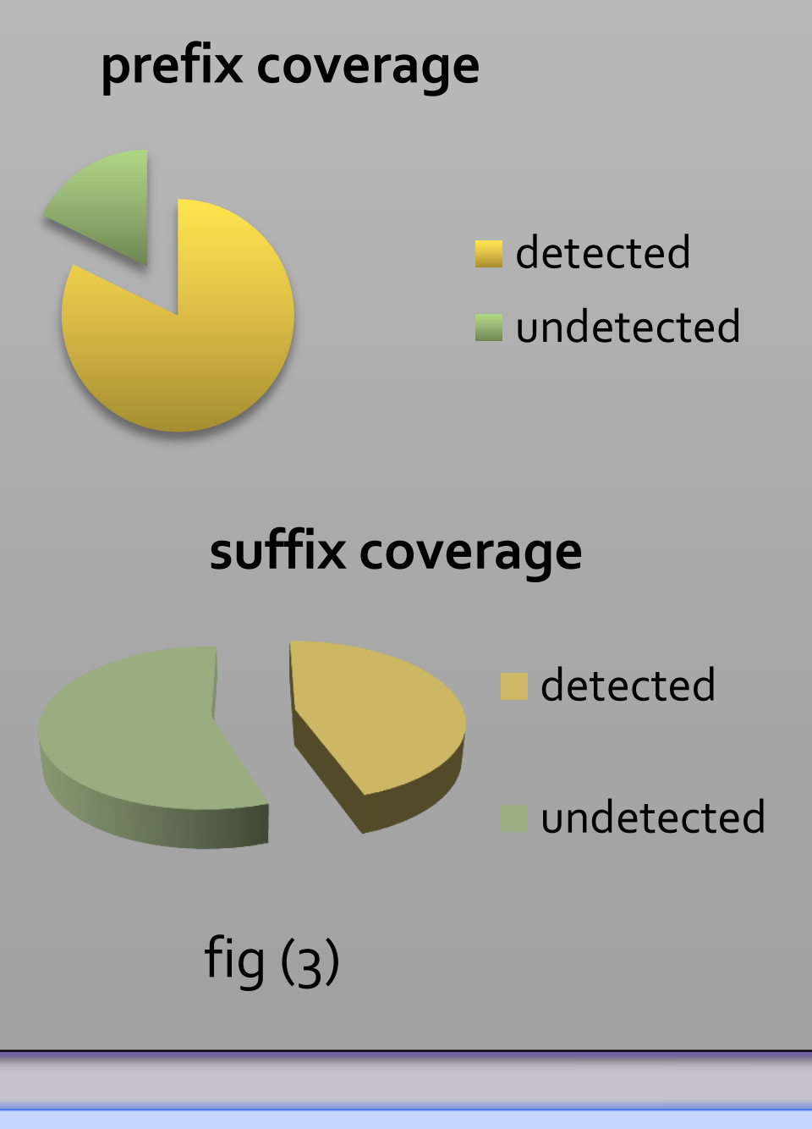


fig (3)

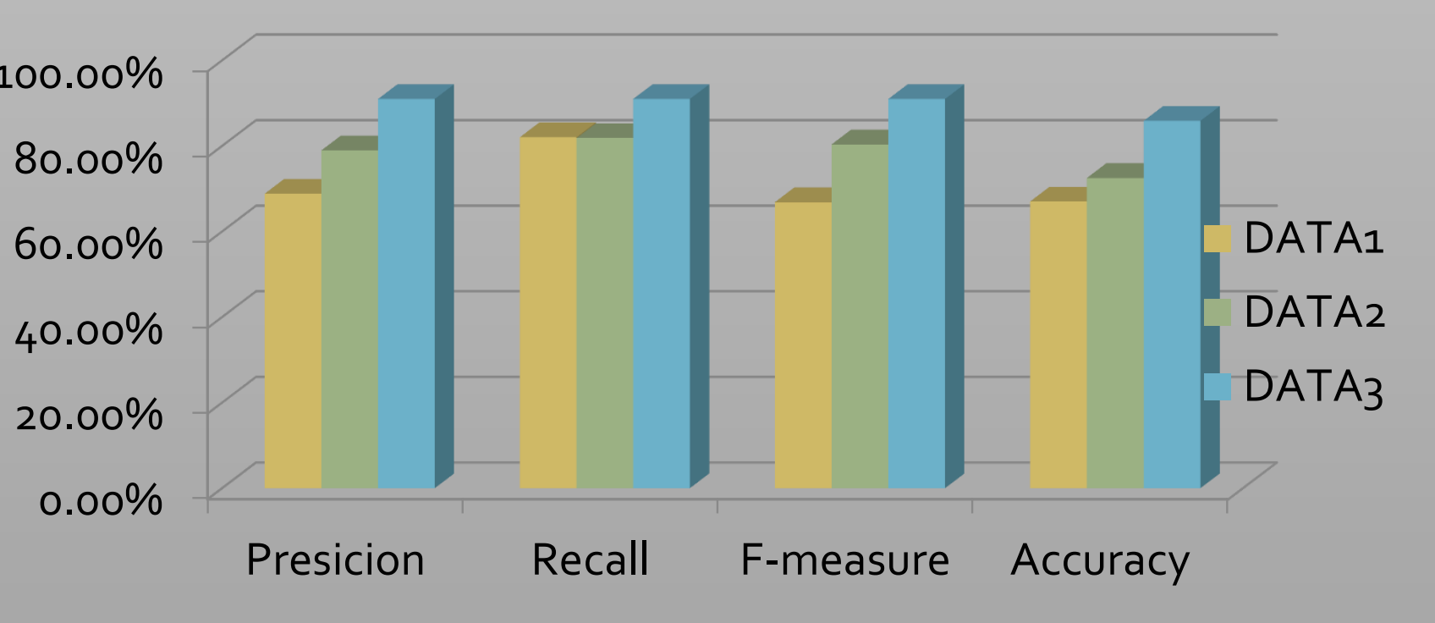


fig (2)

5. CONCLUSION AND DISCUSION

The study shows that the similarity constraints of the Arabic verbal root have significant consequences for phonotactic modeling and speech segmentation
Morphological segmentation based on phonotactics cues was accomplished by the phonotactics modeling
The model succeeded to cover the majority of its required tasks, it needs to be modified to cover extra tasks and to manage words with successive affixes, and words that are not based on verbal root
The experiment
Future work
Increasing the training corpus for more accurate probabilistic estimation
develop the model to perform a higher accuracy with morphological segmentation
Increase the model ability to recognize sequences of morphemes, and Egyptian phonological variation

6. REFERENCES

Adriaan F,(2011), The Induction of Phonotactics for Speech Segmentation, LOT 3512 JK Utrecht Netherlands
Alderete J., Tupper1 P., Stefan A. , (2012), Phonotactic learning without a priori constraints: Arabic root cooccurrence restrictions revisited, Proceedings of the 48th annual meeting of the Chicago Linguistics Society, John Alderete1
Brown F., Peter V. Robert L., (1992),Class-Based n-gram Models of Natural language
Elshafei M.Ali M., automatic segmentation of Arabic speech, King Abdulaziz City of Science and Technology
Frisch S., Broe M., Pierrehumbert J.,(1997),Similarity and phonotactics in Arabic
Fadi Biadsy and Julia Hirschberg, Habash Spoken Arabic Dialect Identification Using Phonotactic Modeling
Holes. 2004. Modern Arabic: Structures, Functions, and Varieties. Georgetown University Press. Revised Edition.
Watson.j.2007, The phonology and morphology, of Arabic, Oxford university press