# Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic

**Emad Mohamed, Behrang Mohit, Kemal Oflazer**

Carnegie Mellon University-Qatar

Doha, Qatar

emohamed@qatar.cmu.edu, behrang@cmu.edu, ko@cs.cmu.edu

**Abstract**

We present an annotation and morphological segmentation scheme for Egyptian Colloquial Arabic (ECA) with which we annotate user-generated content that significantly deviates from the orthographic and grammatical rules of Modern Standard Arabic and thus cannot be processed by the commonly used MSA tools. Using a per letter classification scheme in which each letter is classified as either a segment boundary or not, and using a memory-based classifier, with only word-internal context, prove effective and achieve a 92% exact match accuracy at the word level. The well-known MADA system achieves 81%, while the per letter classification scheme using the ATB achieves 82%. Error analysis shows that the major problem is that of character ambiguity, since the ECA orthography overloads the characters which would otherwise be more specific in MSA, like the differences between y (ي) and Y (ى) and A (ا ) < , (إ), and < (أ) which are collapsed to y (ي) and A (ا) respectively or even totally confused and interchangeable. While normalization helps alleviate orthographic inconsistencies, it aggravates the problem of ambiguity.

**Keywords:** morphological segmentation, Egyptian Arabic, memory-based learning, colloquial Arabic annotation

## 1. Introduction

Egyptian Colloquial Arabic (ECA) is the most widely used dialect of the language and is usually written in informal discourse. ECA deviates from Modern Standard Arabic (MSA) at the lexical, morphological, and syntactic levels. Although there is an abundance of MSA resources, they do not seem to be adequate for handling ECA. We present data and a segmentation scheme for ECA that can be used either for lexical research or as a basis for other NLP tasks including part-of-speech tagging and syntactic parsing. We have annotated a corpus of ECA user-generated comments and jokes, segmented by two annotators with a Kappa score of 99.27%. Moreover, we have built a memory-based segmenter for ECA text. We demonstrate that a simple ECA segmenter outperforms the state-of-the-art MSA segmenter on ECA text. These results are motivating for further investment in the creation of specific datasets and tools for processing colloquial Arabic. We plan to release our annotated corpus to the research community.[1]

## 2. Segmenting Egyptian Arabic

Currently there is a rich set of annotated data and processing tools for MSA. The performance of these resources for robust processing of colloquial Arabic is expected to vary. We start with examining ECA text to characterize the differences between MSA and ECA and their effects on ECA processing. We see that the differences are distributed to different linguistic layers which make processing ECA a hard problem. The following differences illustrate (part of) the problem:

1. The prefix preposition ***l+*** (ل, for; to) is treated as a suffix when preceded by a verb and followed by an object pronoun. For example, the word ***msmEtlhm$*** (مسمعتلهمش, I did not listen to them) comprises a negation circumfix (***m*** … ***$***, م … ش), a perfective verb (***smE***, سمع), a first person singular subject prefix (***t***, ت), a preposition (***l+***, ل), and a third person plural object pronoun (***hm***, هم).[2]
2. Morpheme boundaries assimilate when the last letter of one morpheme is the same as the first letter of the next morpheme, in what in MSA are two separate words. For example, when the verb ***qAl*** (said, قال) is followed by the preposition ***l+*** (to, ل), one of the two ***l***'s disappears: ***qAlhA*** (He said to her, قالها). The missing letter is compensated for by means of consonant gemination, which is hardly discernible in the orthography.
3. The writing system is not standardized and there is plenty of variation. The forms ***qlh*** (قله), ***qAlw*** (قالو), ***qAlwA*** (قالوا), and even sometimes ***AlwA*** (الوا), have been found to represent the same linguistic unit, translated as "*He said to him*" in English.
4. The Arabic glottal stop, known as the ***hamza***, is written in different ways in MSA and the

1 The segmented corpus, as well as supporting scripts and this paper, is available at http://www.qatar.cmu.edu/~emohamed/

2 Throughout this paper, Arabic words are presented in the Buckwalter encoding followed by the word in the Arabic script and an English gloss in parentheses.

situation gets worse in ECA. Most forms of the hamza are interchangeable in the colloquial: ***Al>xwAn*** (الأخوان), which in the standard orthography means *the two brothers* has been used in place of ***Al<xwAn*** (الإخوان), which means *the Muslim Brotherhood.*

5. The singular masculine pronoun **h** (ه) and the word-final feminine marker **p** (ة) are only distinguished by the two dot diacritic on the latter, but are usually interchangeable in the colloquial.
6. Long vowels shorten and disappear in the orthography. There is a tendency in ECA to shorten long vowels before constant clusters and in certain morphological templates (Abdu-Mansur, 1990), in which long vowels turn into their equivalent short ones. Since short vowels are not written in Arabic, we often end up with the vowel missing altogether. An example of this is the verb ***yAxud*** (he takes, ياخد) which turn into ***yaxud*** (يخد) when an object pronoun is cliticized, thus forming a consonant cluster (***yAxudhum*** → ***yaxudhum***). The same also holds true for nouns with possessive clitics. Based on a small sample, this phonetic phenomenon is mostly reflected in the orthography, although this is not uniform due to the non-standardization of the writing rules for colloquial dialects.

These differences and other syntactic characteristics like ECA's word order demonstrate the limits of MSA-based systems and necessitate the creation of new resources and tools for ECA processing. In this paper, we focus on the task of morphological segmentation of ECA.

## 3. Annotation

We trained two undergraduate students who are native Arabic speakers, as annotators. The annotation task is to segment Arabic words in context, using a simple framework: whenever there is a segment boundary, the annotators add a "+" sign between segments. For example for the word ***mfhmthm$*** (I did not understand them, مفهمتهمش***)***, annotators are expected to mark the word as ***m+fhm+t+hm+$*** where the ***m*** is the first part of the negative circumfix, ***fhm*** is the verb, ***t*** the subject suffix, ***hm*** the object pronoun, and **$** the second part of the negative circumfix. The training included learning about segment functions, but the annotation task was limited to only segmentation.

The data comprised user-contributed (political) comments and jokes from the Egyptian web site www.masrawy.com. The users of this web site tend to use colloquial words and structures more often than many other websites, and even the web site editors themselves incorporate a fair amount of colloquialness in their reporting.

We manually selected the colloquial comments and excluded the texts that are more MSA than colloquial. Following the training and some trial annotations, we measured the inter-annotator agreement on a sub-corpus of 2899 words. We observed a Kappa score of 99.27% between the two annotators.

While the training set varies per experiment below, the colloquial training set comprises 320 comments and 20022 words. The test set comprises 36 comments consisting of 2445 words including punctuation. The average number of words per comment is 68.

It is worth noting that the plurality of the words in the training set (45%) are made up of only one segment (roughly a single morpheme). 37% of the words are bi-segmental and 14% are tri-segmental. Only two words have more than 7 segments, and both are typos. We have decided to keep the orthographic errors as is, since our purpose is to model this language variety as written and/or spoken by the users.

## 4. Building a simple ECA segmenter

In order to test the utility of our annotated dataset, we built a simple ECA segmenter and compared its performance with the MSA-based systems. We trained a memory-based learner in a per letter classification task to detect segment boundaries.

Memory-based learning is based on the idea that instances during learning are stored in memory, and when a new instance is encountered, the closest instance in memory is returned based on some distance metric (Daelemans *et al.*, 2010).

| **-5** | **-4** | **-3** | **-2** | **-1** | **0** | **1** | **2** | **3** | **4** | **5** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| _ | _ | _ | _ | _ | **m** | q | l | t | h | l | + |
| _ | _ | _ | _ | m | **q** | l | t | h | l | h | _ |
| _ | _ | _ | m | q | **l** | t | h | l | h | m | + |
| _ | _ | m | q | l | **t** | h | l | h | m | $ | + |
| _ | m | q | l | t | **h** | l | h | m | $ | _ | + |
| M | m | l | t | h | **l** | h | m | $ | _ | _ | + |
| Q | l | t | h | l | **h** | m | $ | _ | _ | _ | _ |
| L | t | h | l | h | **m** | $ | _ | _ | _ | _ | + |
| T | h | l | h | m | **$** | _ | _ | _ | _ | _ | _ |

Table 1: Character-based feature set: The focus character (0) is in bold, the negative numbers indicate the characters preceding the focus character, and the positive ones the characters following the focus character. The last column is the class, which is either + or _

We used the Timbl memory-based learner (Daelemans *et al.*, 2010) with a very basic feature representation in which we used only the preceding five characters and the following five characters, when present, as features (See

Table 1 for feature representation of the word ***m+ql+t+l+hm+$*** (I did not say to them, مقلتلهمش)). We used the Timbl IB1 algorithm with similarity computed as overlap, using weights based on gain ratio, and the number of *k* nearest neighbours equal to 1. These settings were reported to achieve an accuracy of 98.15% when trained and tested on standard Arabic Treebank Data (Mohamed, 2010). These experiments also showed that the wider context and part-of-speech tags, have only a very limited effect on segmentation quality, and that word-internal context alone is enough for producing high quality segmentation. We do not experiment with the sentence context and part-of-speech tagging here and plan to investigate them in the future.

### 4.1 Baseline Segmenters

We compare our ECA segmenter against two baseline MSA segmenters. We ran two baseline experiments both based on using MSA tools to segment ECA:

1. MADA (Habash and Rambow, 2005) which can be viewed as the state-of-the-art MSA segmenter.
2. A memory-based segmenter with the settings above on the Arabic Treebank (ATB p1v3, (Maamouri and Bies, 2004))

We apply normalization throughout this paper, where indicated, in which we replace the ***taa marbuta*** (p, ة) with the ***h*** (ه), the different forms of alif-hamza (<>|, أإآ) with ***alif*** (ا), and the final ***y*** (ي) with ***Y*** (ى). The output of MADA was normalized and compared with the normalized test set, as MADA attempts to restore standard orthography, and we did not want this restoration to be penalized. While normalization helps smooth the data and reduce data sparseness, it also has the negative effect of increasing ambiguity. For example, the word **yAsr** (ياسر) in standard orthography is a proper noun, but when it gets normalized, it can be either ***y>sr*** يأسر (segmented as ***y+>sr***) or ***yAsr***, which consists of one morpheme. Also, while the final ***y*** and ***Y*** have completely different functions, as the latter cannot be a separate segment, unlike the earlier, in normalization, this distinction is lost. For example, the standard word **bSrY** (بصرى) is a city name, but the normalized form could mean the city name, my-eyesight, or optical, with the possessive form having two morphemes.

### 4.2 Experimental setup

We ran five experiments for ECA word segmentation:

1. ***MADA:*** The baseline MSA segmenter.
2. ***MSA:*** Our baseline MSA segmenter (trained on the ATB data).
3. ***ECA:*** Our ECA segmenter that we train on the colloquial training set and test on the test set in their original forms, without applying any normalizations.
4. ***ECA+norm:*** Similar to ECA, with an addition of orthographic normalizations to both the training and test sets
5. **ECA+MSA+norm** in which we normalize the ATAB p1v3 section and add it to the normalized training set.

The results of these experiments are in Tables 2, 3, and 4.

## 5. Results and Discussion

For evaluation, we use word accuracy (exact match). Although we use letters in classification, we count a word as correctly segmented if and only if all the letters in the word are correctly marked. This is a harsh measure and does not provide any credit to many words which are almost correct. However, we believe that correct segmentation influences subsequent tasks such as part-of-speech tagging and syntactic parsing, and that any errors in segmentation have an adverse effect on the subsequent processes.

In spite of the exclusion of letter accuracy as a metric, we still give the numbers in the result tables below as they give an indication of partial accuracy.

We also report on the performance on unknown words, those words in the test set that are not in the training set. This helps us discover how generalizable the results presented here are. Since the training set varies by experiment, the ratio of unknown words varies accordingly.

Table 2 presents the results of the five experiments. We observe that the simple ECA-based segmenter significantly outperforms the state-of-the-art MSA segmenter (MADA). Considering the simplicity of the learning framework and the small feature set, we find this result interesting. The lexical coverage is one of the main reasons for this performance differences. For MADA, we observed that 189 of the 468 (40%) words with segmentation errors were not found in the underlying lexicon and received a NO-ANALYSIS tag. MADA has an accuracy of 88.16% if we exclude the NO-ANALYSIS words. It has to be borne in mind that MADA was developed for MSA and that it is only natural for it to miss the colloquial words.

The accuracy increases with normalization and lexicon expansion (from the ATB data) until it reaches its highest (91.90%) with the ECA+MSA+norm experiment. We also notice that accuracy increases as the percentage of unknown words drops (See table 3 on the accuracy of out of vocabulary words). There is a very strong negative Pearson correlation co-efficient of 0.9885 between the percentage of unknown words and accuracy, which indicates that lexical coverage plays a major role in the segmentation accuracy and that adding more data should improve the results.

The results of known words (or In-Vocabulary words) in table 4 corroborate the proposition that lexical coverage is the key factor in segmentation accuracy. One noticeable result in table 4 is that the accuracy on known words decreases with normalization. This is natural, since normalization leads to more ambiguity, as a single written form may have more than one possible segmentation.

| | Experiment | Word Accuracy (%) | Letter Accuracy (%) |
|---|---|---|---|
| 1 | MADA | 81.48 | - |
| 2 | MSA | 81.52 | 95.11 |
| 3 | ECA | 88.50 | 96.70 |
| 4 | ECA+norm | 89.20 | 96.90 |
| 5 | ECA+MSA+norm | 91.90 | 97.80 |

Table 2: Segmentation experiments and results

| | Experiment | OOV (%) | Word Accuracy (%) |
|---|---|---|---|
| 1 | MADA | - | - |
| 2 | MSA | 59.88 | 69.81 |
| 3 | ECA | 30.59 | 65.37 |
| 4 | ECA+norm | 29.45 | 66.39 |
| 5 | ECA+MSA+norm | 23.68 | 70.81 |

Table 3 : Accuracy on out-of-vocabulary words (OOV) across the segmentation experiments. Column 3 presents the percentage of OOV words in the training set

| | Experiment | IV (%) | Word Accuracy (%) |
|---|---|---|---|
| 1 | MADA | - | - |
| 2 | MSA | 40.12 | 99.00 |
| 3 | ECA | 69.41 | 98.69 |
| 4 | ECA+norm | 70.55 | 98.72 |
| 5 | ECA+MSA+norm | 76.32 | 98.44 |

Table 4 : Accuracy on in-vocabulary words (IV) across the segmentation experiments. Column 3 presents the percentage of in-vocabulary words per training set

## 6. Error Analysis

The word-final letter ***Y*** (ى) alone accounts for 10.6% of those letters that did not get segmented although they had to (false negatives). This is an ambiguous letter as it can be segmented (with different meanings and part-of-speech tags) and can also be a derivational suffix which does not require segmentation. The segment ***nA*** (we; us, نا) ranks second with 8%. In the case of over-segmentation, the letter **h** (Arabic: ه) alone accounts for 28.33% of over-segmentation errors, followed by ***yn*** ين) and **Y** at 5% each.

We also noticed that the colloquial data could lead to errors in the segmentation of standard words. For example, the word **mEtqdAthm** ( معتقداتهم ) is correctly segmented as ***mEtqd+At+hm*** when using only the ATB, but incorrectly as ***m+Etqd+At+hm*** with ECA +MSA+norm. This may be due to the frequency of ***m*** as a prefix in the colloquial training data.

## 7. Conclusions

We have presented an on-going corpus collection process for creating tools and datasets for processing Egyptian colloquial Arabic motivated by the fact that the performance of the tools trained on MSA decays as they are faced with dialectal norms and orthography. We achieved promising segmentation results by combining the dialectal data with MSA data. We plan to invest more in the learning side of morphological segmentation along with the ongoing annotation process. The data and the tools will be released to the research community.

## 8. Acknowledgements

We thank Mariem Fekih Zguir and Afrah Hassan for assistance with annotation. We also thank the anonymous reviewers for their instructive and enriching feedback. This publication was made possible by a NPRP grant (NPRP 09-1140-1-177) from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

## 9. References

Abdu-Mansur, Mahasen Hassan. Vowel Shortening in Two Arabic Dialects. In Broselow, Ellen, Eid, Mushira and McCarthy, John (Editors). Perspectives on Arabic Linguistics: Papers from the Annual Symposium on Arabic Linguistics, Detroit, Michigan 1990 (Current Issues in Linguistic Theory).

Buckwalter, Tim (2002). Arabic morphological analyzer version 1.0. Linguistic Data Consortium. LDC Catalogue Number: LDC2002L49.

Habash, Nizar and Rambow, Owen. 2005. Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop, In Proceedings of ACL'05.

Habash, Nizar, Rambow, Owen and Roth, Ryan. 2010. The MADA and TOKAN Manual.

Maamouri, Mohamed and Bies, Ann (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools, In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, August 28, 2004.

Mohamed, Emad (2010). Orthographic Enrichment for Arabic Grammatical Analysis. PhD Dissertation, Indiana University, USA.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch and Ko van der Sloot . 2010. TiMBL: Tilburg Memory-Based Learner. Reference Guide.