# Machine Translation of Egyptian Arabic

# Using Morphological Analysis

**Weston Feely, Serena Jeblee**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA, 15213
wfeely@cs.cmu.edu, sjeblee@cs.cmu.edu

## Abstract

We present the results of a Machine Translation experiment for Egyptian Arabic to English, using a newly created morphological analyzer for Egyptian Arabic and the freely available Moses decoder. The translation system we create is trained on an Egyptian Arabic-English bitext and a Levantine Arabic-English bitext. The BLEU scores attained on held out test sets for three data conditions (Egyptian data only, Levantine data only, and Egyptian and Levantine data) are compared, for each data set: with no morphological analysis, pre-processed using our own morphological analyzer, and pre-processed using MADA, a freely available morphological analyzer for Modern Standard Arabic. The best BLEU score of 17.51 is obtained using the MADA analyzer on the full Egyptian and Levantine data sets.

## 1 Introduction

While there has been much research done on Modern Standard Arabic (MSA) in the Machine Translation community, there has been little work done on dialectal Arabic. We attempt to address the lack of research in dialectal Arabic MT by building an MT system for Egyptian Colloquial Arabic, the most spoken dialect of Arabic.

MSA is the standardized variety of Arabic used for formal speech and writing, which is based on millenia of Arabic literary history and is taught in schools. However, residents of Arabic-speaking countries may not understand it well if they are not highly educated, especially if they do not need to use it in their daily lives.

MSA exists in diglossia with a number of Arabic dialects. Diglossia is a phenomenon where two languages are used within a society, with one language usually being the formal language, and the other language being the vernacular. In Arabic-speaking countries, MSA is the language used in formal register while the Arabic dialects are used in colloquial discourse. However, the usage of the term "dialect" for Egyptian Arabic is a convention; there is not a high level of mutual intelligibility between MSA and the Arabic dialects. This situation makes language processing difficult because natural language processing tools and MT systems trained on MSA bitexts are not very accurate when used to translate Arabic dialects.

Additionally, there is no standardized spelling for Arabic dialects, because until recently MSA was the only written form of Arabic, and so most of the Arabic corpora available for NLP and MT applications are written entirely in MSA.

However, with the rise of the internet and social media, people have begun to write in dialectal Arabic despite the fact that Arabic dialects have no standardized spelling. In colloquial discourse, native Arabic speakers use their local dialects, many of which have low mutual intelligibility with each other as well as with MSA. Many of these dialects have several million speakers; Egyptian Arabic has 54 million speakers worldwide. Additionally, Egyptian Arabic is widely understood in other Arabic-speaking countries because of Egypt's large entertainment and movie industry.

Although MSA and Egyptian Arabic have a shared historical origin, some

common vocabulary and many of the same verb roots, Egyptian Arabic has morphological and syntactic differences from MSA, as well as many more borrowed words from French and English than MSA. Most of the function words and common expressions in Egyptian Arabic are completely different from MSA function words and expressions.

Figure 1 is an example of these morphosyntactic differences between MSA and Egyptian Arabic. The sentence below means "I do not speak Arabic."

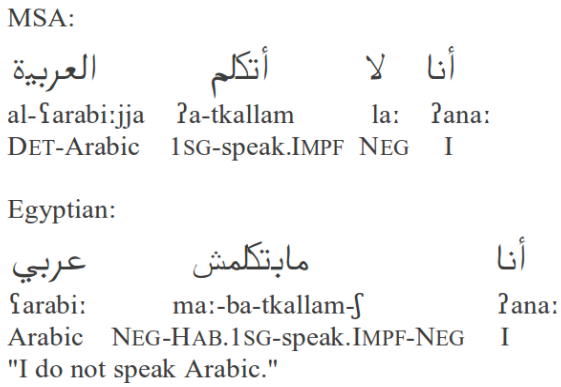


Figure 1: MSA and Egyptian Arabic Morphology Example

This is an example of how MSA and Egyptian Arabic differ in their use of morphological affixes. The same verb for "speak" is accompanied by a negation auxiliary "la:" in MSA, while in Egyptian Arabic, a negation circumfix "ma- -ʃ" performs the same grammatical function. The Egyptian Arabic verb prefix "ba-" carries the first person singular agreement features for the verb, but also carries habitual aspect. In contrast, the MSA first person singular verb prefix "ʔa-" does not have habitual aspect, since MSA does not have a habitual aspect marker.

In order to correctly translate such Egyptian Arabic verbs, these complex morphological affixes must be segmented and translated separately. However, currently there is no freely-available morphological analyzer that can do this for Egyptian Arabic and popular morphological analyzers for MSA will fail to capture the morphemes like the negation circumfix described above, when applied to Egyptian Arabic, which differ drastically from MSA morphemes.

As an example of current state-of-the-art MT for MSA and Egyptian Arabic, below is the output of the online Google Translate system for Arabic, for the same sentences as in the above figure. The Google Translate output for the above MSA sentence is the correct translation:

MSA: I do not speak Arabic
EGY: I Mapetklmh English

Figure 2: Google output of Arabic sentences from figure 1

Google Translate performs significantly better on MSA, as expected, because it is trained on MSA data. Figure 2 serves as an example that translation for Egyptian Arabic cannot be accomplished by a system that was only trained on MSA data, which demonstrates the need for Egyptian Arabic-English MT.

## 2 Previous Work

We base our approach to Egyptian Arabic-English MT on Zbib et al (2012). The data sets for our experiments were also provided by Zbib et al, which included an Egyptian Arabic-English bitext and a Levantine Arabic-English bitext. In their paper, Zbib et al build an MT system for Egyptian Arabic and Levantine Arabic, and they demonstrate that morphological decomposition using MADA (Habash and Rambow 2005) improves Arabic translation, although the morphological analyzer MADA is designed for MSA and used for dialectal Arabic.

Zbib et al (2012) acquired from the LDC monolingual Arabic data that had been crawled from the web. They created their bitext corpora by hiring workers on Amazon Mechanical Turk to identify the dialect of each sentence and translate it into English. They also attempted to filter out segments that were very similar to MSA, and they also attempted to filter junk sentences from the resulting bitext.

Despite the quality control measures that Zbib et al used, the data was still translated by amateurs and therefore the English reference translations are not always fluent or correct. It is also important to note that the Arabic data was pulled from the web, and is also not necessarily completely grammatical, like an MSA-English bitext would be. Additionally, as mentioned before, there is no standardized spelling for Arabic dialects, so the spellings across the bitexts are not consistent.

Zbib et al created different datasets by combining bitexts for each of the different dialects, and combining this dialectal data with large amounts of MSA data. Their results show that combining dialects results in an improvement in BLEU scores on a held out test set, because Levantine Arabic and Egyptian Arabic have some similarities, and each the Egyptian dialectal bitext was quite small on its own. Also, adding MSA to the target dialect can improve translation simply because the training set is so much larger.

## 3 Methods

Our task is to build an Egyptian Arabic to English MT system, using the data sets provided by Zbib et al. Our experimental design includes using Moses (Koehn et al, 2007), an open source decoder, to train a phrase-based machine translation system on several data sets. The three data sets are the Egyptian Arabic-English bitext, the Levantine Arabic-English bitext, and the two bitexts combined. We randomly choose 80% of the sentence pairs in each bitext as a training set, 10% as a development set for tuning, and the remaining 10% are used as the test set for each language condition. Table 1 shows the size of each dataset.

| Data Set | Train | Dev | Test |
|---|---|---|---|
| EGY | 260K | 36K | 36K |
| LEV | 916K | 116K | 116K |
| EGY+LEV | 1.2M | 152K | 151K |

Table 1: Data set sizes in number of words

We train three translation systems, one on each data set, using Moses (Koehn et al, 2007), with the only modification to the data being text normalization, a common pre-processing step for consistency in Arabic text languages, which removes Arabic diacritics from the text and normalizes the letter "alif" and final form of the letter "ya". We report the BLEU scores for these three systems as our "no analysis" condition, for each language condition's test set, respectively.

We then train three more translation systems using the same data, with morphological analysis performed on the data beforehand, using MADA, a freely available MSA morphological analyzer and tokenizer created at Columbia University by Habash and Rambow (2005). MADA was created using a large MSA treebank, and it creates multiple morphological analyses for each word and chooses one best analysis using an SVM classifier trained on morphosyntactic features. It then uses those analyses to tokenize the text, separating off prefixes and suffixes. The resulting three BLEU scores of the MT systems trained on MADA-analyzed data for the MADA analyzed test sets for each language are used for comparison with our own morphological analyzer.

We built our own Egyptian morphological analyzer using FOMA (Hulden, 2009), a freely available finite state transducer-based morphological analysis toolkit. We chose FOMA because it allows for rapid development of a morphological analyzer and it supports Arabic text.

The analyzer matches different parts of speech and various prefixes and suffixes that words can take. Among the suffixes we split off from their stem were pronominal enclitics, which specify the object of a verb, as well as... The prefixes we split off from their stems included prepositions, future and habitual markers, in addition to determiners. Finally, we split off the negation circumfix. We established a level of segmentation that produced tokens which would align well with the English tokens, based on an analysis of the text, which included leaving number and gender markers attached to the stem. The output of our morphological analyzer is a segmentation of the input word, which may be several new pseudo-words. Figure 3 shows the kind of morphology-based segmentation that our analyzer produces.

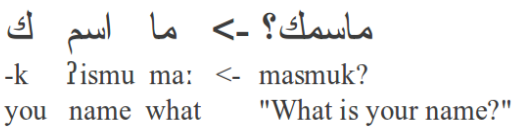


Figure 3: Morphological decomposition

This figure demonstrates the segmentation of prefixes; in this case it is the question marker "ma:" which attaches to the noun in this case. Here, "-k" is a pronominal enclitic indicating that the noun is genitive and "you" is the owner of the noun. While this question, "what is your name?" is four words in English, it is only one in Arabic (as seen in the sentence on the right). This segmentation produces three words (left), which provides more tokens that can align to the English words.

Our FOMA analyzer produces multiple analyses for some Egyptian Arabic words, in which case we randomly choose one analysis, since we did not have labeled data to train a classifier to choose the best analysis, as MADA does. We pre-process the original data from the "no analysis" condition using own FOMA analyzer, and report the BLEU scores on the FOMA analyzed test sets for each language condition.

Additionally, we used FOMA to create Levantine-to-Egyptian Arabic transformation rules, which replace single Levantine lexical items with their Egyptian Arabic equivalents. Although Levantine is classified as a different dialect, it is perhaps the most similar dialect to Egyptian, and the syntax of both dialects is similar. By converting certain function words and morphological patterns to be more like Egyptian, we leverage the larger amount of Levantine data that is available to train a better Egyptian translation system. Figure 4 is an example of one of the lexical substitutions that the transformation produces.

شلون اختك؟ <- ازي اختك ؟

ʔoxtak ʔizajj <- ʔoxtak ʃlo:n

"How is your sister?"

Figure 4: Levantine to Egyptian transformation

The two sentences in this example have the exact same meaning in English, but the first one (right) is typical of Levantine Arabic, while the second one (left) is typical of Egyptian Arabic. We want to use the Levantine data to improve translation for Egyptian, so we can easily convert this Levantine sentence into an Egyptian sentence by replacing the Levantine word for "how" with the Egyptian word for "how."

In order to determine whether these transformations rules have a greater effect than the segmentation from our morphological analyzer on our BLEU score improvements over the "no-analysis" condition, we separate out the analysis and transformation into two FOMA modules. The first module was our FOMA morphological analyzer with no lexical replacement rules. The second module was our FOMA transformation rules with no morphological analysis. We used each module pre-process the original data for three additional conditions: one for the FOMA analyzer only and no transformation rules, one condition for the FOMA transformation rules only and no analysis, and one combined condition for both FOMA analysis and Levantine-to-Egyptian transformation rules. Due to lack of time, we only report the BLEU scores for the Egyptian and Levantine combined data set condition, in the separated FOMA analysis only and Levantine-to-Egyptian transformation rules only conditions.

## 4 Results

| Data Set | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| EGY | 13.26 | 15.28 | | N/A | 13.88 |
| LEV | 15.80 | 17.10 | | | 16.03 |
| EGY+ LEV | 16.09 | 17.51 | 16.35 | 16.17 | 16.39 |

Table 2: BLEU Scores Results from Moses Decoder
1. No analysis
2. MADA analysis
3. FOMA analysis
4. Levantine to Egyptian Transformation rules
5. FOMA analysis + Levantine to Egyptian transformation rules

In the above table, EGY represents the Egyptian Arabic-English bitext data set, LEV represents the Levantine Arabic-English bitext data set, and EGY+LEV represents the combination of these two data sets. We evaluate the output of our system with BLEU, following Zbib et al, and our BLEU scores are taken from an evaluation on a separate, randomly selected 10% of each data set, which was the test set for each condition. The columns are the different analysis conditions, which are explained in the notes below the table. Each combination of language data set and analysis condition is shown in the table, although the FOMA morphological analysis condition and Levantine-to-Egyptian Arabic transformation rules only condition are not included for the EGY and LEV data sets, due to time constraints. The EGY data set with the Levantine-to-Egyptian transformation only condition is marked "N/A", because this data set contains no Levantine sentences.

## 5 Discussion

We obtain the best result of 17.51 by using the MADA morphological analyzer to pre-process the EGY and LEV data sets, and then

training Moses on these data sets together. The translation systems trained on the EGY+LEV data set perform the best overall because these conditions have the largest amount of data available during training. Likewise, the translation systems trained on the EGY data set performed the worst, because this data set was the smallest.

We treated our "no analysis" condition as the baseline to compare the other results against. Using the MADA analyzer to pre-process the data we had a 2.02 BLEU point improvement on the EGY data set, a 1.3 point improvement on the LEV data set, and a 1.42 point improvement on the combined EGY+LEV data set. Using our own FOMA analyzer to pre-process the data we had a 0.62 point improvement on the EGY data set, a 0.23 point improvement on the LEV data set, and a 0.3 point improvement on the combined EGY+LEV data set.

Overall, the translations systems trained on the MADA-analyzed data performed better than our analyzer on each data set. We attribute this to our random selection of a single analysis, when there are multiple analyses for a single word. In contrast, MADA uses an SVM classifier to choose the best analysis, which performed much better than our random selection. Additionally, MADA was designed with a much larger possible of analyses than our FOMA analyzer currently allows for.

We also report the BLEU scores for the EGY+LEV combined data set, with only our FOMA morphological analysis of the data, and only our Levantine-to-Egyptian Arabic transformation rules applied to the data, in order to distinguish the effect of each component on the resulting translation system. By applying the Levantine transformation rules to the data, we gain only a 0.08 BLEU improvement over the baseline "no analysis" score for this data set. By applying our FOMA morphological analysis alone we gain 0.26 BLEU, and applying both analysis and transformation together gains 0.30 BLEU points over the baseline. This demonstrates that the FOMA morphological analysis resulted in a greater improvement than our lexical replacement rules on this data set, although, even the combined effect of both components was less than the improvement gained by pre-processing the same data with MADA, which gets a 1.42 BLEU improvement.

## 6 Conclusions and Future Work

In the future we would like to test different levels of morphological analysis to find the optimal segmentation scheme for Egyptian Arabic. We currently use a scheme that attempts to create tokens that will align well with English, however, this may not be the best approach, so we would like to empirically determine the best level of segmentation for this translation task.

We would also like to add more complicated rules for converting Levantine into Egyptian. Our current approach only includes some simple lexical replacement rules, but we would like to add word re-ordering and more specific pattern-matching to make the syntax and style of the Levantine sentences more like Egyptian. After this, we would be able to treat the Levantine data as a pseudo-Egyptian, provided that we were able to form a more complete transformation of the Levantine data.

We plan to further investigate the effect of analysis versus transformation rules on the translation results in order to find the best combination of Levantine-to-Egyptian dialect transformation and morphological analysis for translation into English. This will include filling in the missing cells in Table 2, including the morphological analysis only condition and Levantine-to-Egyptian transformation only condition for the EGY and LEV data sets.

In order to the pick the best analysis from the multiple options output by the analyzer, we would like to use an unsupervised machine learning approach to learn to pick the best analysis. This will make our morphological decomposition more accurate, and hopefully improve translation.

To further compare to the results of Zbib et al, we hope to train a system on a combination of MSA and dialect data to see if the addition of MSA data will still improve translation with a dialect-specific morphological analyzer. This is a practical approach because of the volume of monolingual and bitext MSA corpora that are available.

## References

Nizar Habash and Owen Rambow. 2005. *Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop.* In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). Ann Arbor, Michigan.

Hulden, M. 2009. *foma: a finite-state compiler and library*. In EACL 2009 Proceedings, pages 29-32.

Philipp Koehn et al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session. Prague, Czech Republic.

Rabih Zbib et al. 2012. *Machine Translation for Arabic Dialects.* In Proceedings of NAACL 2012. Montreal, Canada.