LARGE-SCALE ELECTRONIC CORPORA AND THE STUDY OF MIDDLE AND MIXED ARABIC

Yonatan BELINKOV
Technion – Israel Institute of Technology, Haifa, Israel

1. *Introduction*

Corpora have been a fundamental component in the study of Arabic from the very beginning: already the Medieval Arab grammarians relied on a rather fixed corpus consisting of the Quran, pre-Islamic poetry, and utterances by Bedouins who were thought to preserve the pure Arabic language.[1] They produced grammatical treatises and large vocabularies from the early centuries of Islam onward. In Europe, the first Latin-Arabic glossaries appeared in the 13th century, with grammatical treatises of Classical Arabic appearing mostly in the 19th century.[2] The Arabic dialects have been studied in the west at least since the 19th century, with a grammar of Moroccan Arabic published already in 1800.[3] The two language varieties – Classical and dialectal Arabic – are often described as situated at the two ends of a continuum.[4] In between we find intermediate, or mixed varieties. In the 20th century, grammars of Middle Arabic have appeared, most notably in the studies by Joshua Blau on mixed texts from the medieval period. Recent years have seen a new trend of studying all types of Middle and Mixed Arabic (MMA) in a common framework – whether medieval or modern, spoken or written – as advocated in previous AIMA conferences.[5] What is common to all studies on the various varieties of Arabic is their reliance on collections of texts, in written or recorded form. Such

[1] Versteegh 2006, p. 57-59, 75.

[2] Versteegh 2006, p. 2, 90.

[3] Versteegh 2006, p. 6, 132; Aguadé 2008, p. 287.

[4] A situation referred to as continuglossia (Hary 2003; Hary 2009, p. 37, 40-44).

[5] There has been some debate on the definition of Middle and Mixed Arabic. One suggestion is to reserve Middle Arabic for *written* texts, both modern and pre-modern, and Mixed Arabic for *spoken*, contemporary, mixed language (den Heijer 2012a, 8; Mejdell 2012, p. 235). I will not subscribe to any one definition and instead will usually refer to Middle and Mixed Arabic together as any Arabic variety with a mixed character.

collections have traditionally been compiled by individual scholars through a tedious, manual process, and are therefore usually limited in size, scope, and availability.

In recent years, some large-scale, electronic Arabic corpora have been compiled and made available to the research community. Most of these corpora have been gathered with a restriction to a certain variety in mind, typically Modern Standard Arabic (MSA). However, in practice there is never a clear distinction between the different varieties, and even text collections that are thought to be purely MSA may include mixed linguistic forms.[6] With the emergence of electronic media, mixed language attestations are becoming more and more common in writing, a development that has been recognized by MMA researchers.[7] Furthermore, the great potential of large-scale corpora for MMA research has been acknowledged in previous AIMA conferences.[8] This paper aims to make a modest step in this direction and advocate the use of large-scale electronic corpora when studying mixed language phenomena. It first surveys a number of Arabic corpora with a particular focus on their potential contribution to MMA research. From Classical to dialectal Arabic, via Judeo-Arabic and code-switched data, major collections of texts are highlighted in the context of MMA. Then, an automatic, statistical method for identifying language varieties is presented, along with an application to the identification of mixed texts. This is followed by a case study of applying the method to a particular large corpus of web texts. Finally, the concluding remarks offer some directions for future research.

---

[6] For example, even the Gigaword corpus (on which see below), allegedly a pure-MSA corpus, contains a non-negligible number of dialectal words (Mubarak and Darwish 2014).

[7] Mejdell 2012, p. 244-245.

[8] Den Heijer 2012a, p. 20-21; Mejdell 2012, p. 239. A related need is the development of databases of linguistic features, highlighted by Lentin, Grand'Henry (as cited in den Heijer 2012b). While electronic corpora and databases can be mutually beneficial, this topic is beyond the scope of the present paper.

2. *Survey*

Large-scale Arabic corpora available in electronic format are relatively scarce compared to other languages. However, a rising interest in Arabic in the natural language processing (NLP) community has led to the development of many useful electronic corpora. Most of these are primarily comprised of (reportedly) MSA texts, but more and more collections of texts in other Arabic varieties are becoming available. This survey aims to highlight a number of large-scale electronic Arabic corpora relevant to the study of MMA.

Before describing available text corpora, a note is in order about speech corpora, i.e., recordings of spoken Arabic. Such recordings play a crucial role in many studies of spoken Arabic, especially in dialectological research. Their importance to MMA has also been recognized (Mejdell 2012). However, such corpora are often of limited size and are usually not publicly available. A notable exception is SemArch, the archive of Semitic languages at the University of Heidelberg.[9] This archive contains some 2,000 audio recordings (as of January 2016), many of which are in various Arabic dialects. On the other hand, most of the recordings do not have associated transcriptions, instead referring to printed publications. Available transcriptions are typically in PDF format and are not readily available for download and processing by automatic methods.

Other, more accessible speech corpora have been produced and used by the speech recognition community.[10] Some of these are quite large, ranging from dozens to hundreds of hours of recorded speech. Many come with accompanying transcriptions. They cover a number of national Arabic dialects as well as MSA. Such resources are usually produced by organizations and companies like LDC, ELRA, and Appen,[11] and are often not freely available. More importantly, however, these speech corpora are usually designed with the purpose of building speech recognition applications and not for linguistic analysis. One implication of this is that their transcription style does not follow

[9] http://www.semarch.uni-hd.de.

[10] A fairly recent list of Arabic speech corpora is found in Habash 2010, p. 133-134.

[11] https://www.ldc.upenn.edu, http://www.elra.info/en and http://www.appen.com.

standard practices in Arabic linguistics. For example, it often employs the Arabic script for transcribing dialectal speech, with all its known orthographic limitations. Any study relying on such resources will have to take this limitation under consideration. Having briefly noted the potential contribution of large-scale speech corpora to MMA, the rest of this survey focuses instead on written text corpora.

Several previous surveys have outlined different aspects of available electronic Arabic text corpora. Perhaps the first such survey is given by Al-Sulaiti, whose purpose was to compile a "corpus of contemporary Arabic, which would include not only Standard Arabic but also samples of colloquial varieties".[12] Her corpus is in itself a diverse corpus containing about 800,000 words from various genres, including children's stories, interviews, politics, and recipes.[13] Despite having been compiled more than a decade ago, this survey includes useful references to certain Arabic corpora that are otherwise often neglected. Such, for example, are the Buckwalter corpus,[14] the Leuven corpus,[15] and the Nijmegen corpus.[16] These corpora were mostly compiled for lexicographic purposes and do not appear to be readily available. More well known are ELRA's An-Nahar Newspaper Text Corpus[17] and LDC's Arabic Gigaword corpus (Parker et al. 2011), both of which are predominantly MSA news texts.[18] The latter, in particular, has been continuously updated and its 5th edition includes over 1 billion words in more than 3 million documents. It is by now a common corpus in Arabic NLP research.[19]

---

[12] Al-Sulaiti 2004, p. i, 5-15.

[13] http://www.comp.leeds.ac.uk/eric/latifa/research.htm.

[14] http://www.qamus.org/wordlist.htm

[15] http://ilt.kuleuven.be/arabic/ENG/indexENG.php.

[16] An archived version is found here: https://web.archive.org/web/20120122061356/http://www.let.kun.nl/wba/Content2/1.4.5_Nijmegen_Corpus.htm

[17] ELRA catalogue number ELRA-W0027.

[18] Although it contains a non-negligible number of dialectal words; see note 6 above.

[19] Dozens if not hundreds of studies referencing the Arabic Gigaword are found on the ACL Anthology (http://www.aclweb.org/anthology) and on Google Scholar.

What is evident from Al-Sulaiti's survey is that at the time the majority of Arabic corpora were focused on MSA and mostly limited to the news domain. Fortunately, this situation has changed in a number of ways, as is evident by more recent surveys.[20] On the one hand, several corpora aiming for diversity in genre and domain have become available. An example is the Open Source Arabic Corpus (Saad and Ashour 2010), with 18 million words from diverse genres such as sports, stories, and recipes,[21] or the older CLARA corpus (Zemánek 2001), though its availability is unclear. Such genres are more likely to contain mixed forms than traditional news corpora. On the other hand, multiple corpora focused on non-MSA varieties have appeared, including dialectal Arabic and Classical Arabic. These are especially important for MMA research. In fact, to paraphrase Lentin's (2008) words on Classical Arabic texts, one might argue that the language of any *written* dialectal Arabic text should be assigned to some extent to MMA. Below I first survey dialectal Arabic corpora and then move to Classical Arabic corpora.

Especially interesting are so-called multi-dialectal corpora. Often, these are just collections of texts in several dialects coming from similar sources. For example, Almeman and Lee (2013) describe a corpus of Gulf, North African, Levantine, and Egyptian dialects. They collected their corpus by compiling a small list of highly dialectal words in each dialect, as judged by native speakers, and searching for those words with a web search engine. The entire corpus contains around 50 million words and is available by contacting the authors. Another multi-dialectal corpus, compiled by Cotterell and Callison-Burch (2014), includes texts from five regions (Maghreb, Egypt, the Levant, Iraq, and the Gulf) with different dialectal backgrounds, taken from two sources: online newspaper user comments and Twitter tweets. All texts have been human annotated on Amazon's Mechanical Turk and judged as having high dialectal content. The total size is roughly 1.2 million

[20] See Shoufan and Al-Ameri (2015, p. 38-39), for dialectal corpora, and Zaghouni (2014), for freely available corpora, whose list is also maintained at http://www.qatar.-cmu.edu/~wajdiz/corpora.html. Another survey by Al-Thubaity (2015) draws a useful comparison of Arabic corpora in terms of size, domain, medium, and availability.

[21] https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/.

words and it is freely available.[22] Online user comments are also the source of the Arabic Online Commentary by Zaidan and Callison-Burch (2011), a 52-million-word corpus rich in dialectal content. It covers Levantine, Gulf, and Egyptian dialects, and a portion of it was annotated at the sentence level on Amazon's Mechanical Turk.[23] Another multi-dialectal corpus, based on Twitter, is by Mubarak and Darwish (2014), who filtered tweets based on geographical information and highly dialectal words. Their final corpus contains about 6.5 million tweets. Missing from recent surveys, YouDACC is a large corpus of user comments to YouTube videos in Gulf, Egyptian, Iraqi, Maghrebi, and Levantine dialects, with more than 6 million words (Salama et al. 2014). The authors used the user location and a list of keywords to determine the used dialect on the sentence level. Finally, Sadat et al. (2014) collected texts from web forums from 18 countries amounting to over 600,000 words.[24] However, they do not provide many details about their corpus and its availability is unclear.

More limited in scope are several corpora providing parallel texts in several language varieties, i.e. sentences in one variety translated to one or more other varieties. For example, Bouamor et al. (2014) describe a parallel corpus of 2,000 sentences in MSA and several dialects (Eyptian, Tunisian, Jordanian, Palestinian, and Syrian). Meftouh et al. (2015) present a parallel corpus of 6,400 sentences in MSA and several North-African dialects (specifically, Annaba, Algiers, and Sfax), as well as Syrian and Palestinian dialects. This corpus is particularly interesting because it includes dialects at a fine-grained level and is not limited to the regional level. However, both of these parallel corpora are written in Arabic script, which hides many of the dialectal differences.[25]

---

[22] https://github.com/ryancotterell/arabic_dialect_annotation.

[23] The AOC is available at https://github.com/sjeblee/AOC.

[24] These are: Algeria, Bahrain, Egypt, Emirates, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, and Tunisia.

[25] Mejdell (2012, p. 237) mentions the problematic issue of writing conventions in Arabic script.

There have been some efforts to compile corpora for code switching between MSA and Arabic dialects. Such a corpus was provided in the first workshop on computational approaches to code switching (Solorio et al. 2014).[26] It contains roughly 10,000 tweets and 7,000 user comments (overlapping with the AOC mentioned above), in both Egyptian Arabic and MSA. Somewhat ambitiously, the texts were human annotated at the word level as dialect, MSA, mixed, ambiguous, or other. Naturally, such a corpus can be invaluable for MMA research.

In addition to multi-dialectal corpora, a number of mono-dialectal corpora have appeared in recent years.[27] Especially important are so-called "treebanks". These are text corpora with syntactic annotations, or trees. Following the footsteps of the Penn Treebank for English (Marcus et al. 1993) and the Penn Arabic Treebank for MSA (Maamouri et al. 2004), a treebank for Egyptian Arabic has recently been published (Maamouri et al. 2014). Containing more than 300,000 words, it is annotated with morphological and syntactic information. A smaller treebank for transcribed spoken Levantine Arabic also exists, containing some 26,000 words (Maamouri et al. 2006). Both the Egyptian and the Levantine treebanks are available by contacting the LDC.[28] Treebanks of dialectal texts can be especially important for MMA studies thanks to the linguistic annotations they include. Contrary to most large-scale corpora, where automatic and quantitative analyses are limited to looking at surface forms, annotations in treebanks facilitate such analyses on deeper linguistic levels, from morphology to syntax.

From this short survey it should be evident that (written) dialectal Arabic is represented quite well in large-scale electronic corpora. On the other end of the continuum stands Classical Arabic. Although Middle Arabic is often described as neither dialectal nor

---

[26] The dataset is available at http://emnlp2014.org/workshops/CodeSwitch/call.html.

[27] A survey is found in Shoufan and Al-Ameri 2015.

[28] LDC catalog numbers LDC2005E78, LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21. While finalizing this paper, a treebank of Arabic weblog texts appeared, comprising over 200,000 words; see catalog number LDC2016T02.

classical, Lentin (2008, p. 216) argues that the language of Classical Arabic texts should be assigned to Middle Arabic, at least to some extent. It is therefore important to consider available electronic Classical Arabic corpora, even though extracting the unstandardized parts from these texts can be quite difficult.

Perhaps the first large electronic corpus of Classical Arabic is the one compiled by Elewa (2004), comprising 5 million words. It contains texts from the first few centuries in the Islamic era, downloaded from two websites.[29] While appearing to be quite diverse in terms of genres, it is not publicly available and is by now outdated by larger corpora. In particular, the King Saud University Corpus of Classical Arabic (Alrabiah et al. 2013) contains 50 million words from a similar time frame, covering various genres (religion, linguistics, literature, science, sociology, and biography).[30] Despite its limitation to the first four centuries in the Islamic era, it contains texts from genres known to reflect Middle Arabic use (Lentin, 2008, p. 216). Also worth mentioning are a 2.5 million word corpus of "classical Islamic literature" compiled by Rashwan et al. (2011), and Tashkeela,[31] comprising 6 million words of Classical Arabic texts taken from the Al-Maktaba Al-Shamela website.[32] In fact, Al-Maktaba Al-Shamela and similar websites[33] contain numerous books in Classical Arabic that may be valuable for MMA studies.[34] However, most are not easily accessible. In this regard, ongoing efforts to provide an accessible

[29] http://www.muhaddith.org and http://www.alwaraq.com.

[30] The KSUCCA can be downloaded at https://sourceforge.net/projects/ksucca-corpus/.

[31] https://sourceforge.net/projects/tashkeela/.

[32] http://shamela.ws.

[33] For example, http://www.almeshkat.net and http://www.al-eman.com/Islamlib, also mentioned by Habash (2010, p. 135).

[34] For example, a modern commentary to Tafsīr al-Ǧalālayn by ʿAbd al-Karīm al-Xuḍayr (available at https://shamela.ws/index.php/book/23842) contains a few instances of the dialectal phrase *wu-midrīʾēš* ("and I don't know what") inside standard Arabic paragraphs. I thank Alexander Magidow for pointing me to this example.

version of Al-Maktaba Al-Shamela might prove valuable in the future (Belinkov et al. 2016).[35]

Finally, several websites provide online access to large and quite diverse corpora. Even though the actual texts are not available for download, their search interfaces can be valuable for MMA research, although many of the texts on such websites suffer from problems of editing and typing that often hide Middle Arabic norms.[36] Of the web-based corpora mentioned by Zaghouni (2014), the following are especially relevant for our purposes. The KACST Arabic Corpus (Al-Thubaity 2015) contains more than 700 million words from various sources, including around 16 million words from the Internet and roughly the same number of words from texts from the first four centuries in the Islamic era.[37] The Leeds Arabic Internet Corpus[38] has more than 300 million words and the International Corpus of Arabic[39] comprises 100 million words. The latter aims to be diverse in terms of both genre and geography. Even though it purports to be MSA, it includes some genres that are likely to contain mixed language such as novels and plays (Alansary et al. 2007). The well-known ArabiCorpus[40] contains more than 170 million words from diverse sources: from Quran and Hadith, via pre-modern Adab texts like *Kitāb al-ʾaġānī* to modern news and literature. It is therefore an invaluable resource for MMA research. Other web-based corpora, not in Zaghouni's survey, include the Tunisian Arabic Corpus[41] that contains more than 800,000 words, from diverse sources such as folktales, screenplays, web forums,

---

[35] Much work has been done by Maxim Romanov; see also his PhD dissertation (Romanov 2013). Work continues as part of the Open Islamicate Texts Initiative (OpenITI): http://iti-corpus.github.io.

[36] This issue has been recognized in the introductions to both AIMA 1 (by Lentin and Grand'Henry, as cited in den Heijer 2012a) and AIMA 2 (den Heijer 2012a).

[37] See the statistics archived at https://web.archive.org/web/20150905015724/http://www.kacstac.org.sa/Pages/ByMedium.aspx.

[38] The link given by Zaghouni, http://smlc09.leeds.ac.uk/query-ar.html, is not active as of January 2016; an alternate one is at http://corpus.leeds.ac.uk/internet.html.

[39] http://www.bibalex.org/ica/en/About.aspx.

[40] http://arabicorpus.byu.edu.

[41] http://www.tunisiya.org.

and transcribed recordings. It also provides a useful and flexible search interface. Also worth mentioning here is arTenTen (Arts et al. 2014), a 5.8 billion words corpus crawled from the web, with a sub-corpus of 115 million words available through the Sketch Engine (Kilgarriff et al. 2004) query tool.[42]

Before concluding this section, a note is in order about another web-based corpus, the Judeo-Arabic Corpus, maintained by the Friedberg Jewish Manuscript Society.[43] It contains close to 4 million words in 110 Judeo-Arabic texts from the 8th to the 18th century (as of April 2018). Texts can be viewed in both the original scanned image and a transcription. Even though it is not available for download, the web interface allows for quite sophisticated searches for words and phrases. In short, a corpus dedicated for Judeo-Arabic, which may be recognized as a special variety of Middle Arabic (Lentin 2008, p. 218) and a notable example of a mixed language (Hary 2009, p. 29, 37-41; Hary 2012), is an invaluable resource for MMA studies.

3. *Identifying mixed texts*

The previous section presented a number of collections of texts with the potential to contribute to MMA studies. While some of these corpora contain multiple texts in different varieties, others contain individual texts that are likely to be of a mixed character. One potential use case for such corpora is searching them for certain words or phrases that are under investigation. However, if one wants to examine mixed texts without conditioning on specific key phrases, it is not clear how to extract such texts from very large collections. In other words, one might look for an *automatic* computational method for identifying mixed texts. This section describes one such method, based on statistical language modeling. It also investigates its application in the context of a specific corpus and demonstrates its potential contribution to MMA studies.

The approach taken here for automatic identification of mixed texts treats it as a case of language identification. In particular, we are interested in Arabic *variety* identification, that is, identifying whether

[42] https://www.sketchengine.co.uk.

[43] https://fjms.genizah.org.

the variety of a given text is MSA, Classical Arabic, dialectal Arabic, MMA, and so on. Identifying the language of a given text is considered as a relatively simple task in NLP, especially when the languages of concern are distant and the texts are long enough. However, language identification in short texts is still a challenge (Zaidan and Challison-Burch 2014). More significantly, distinguishing closely related language varieties is a non-trivial task that has received recent attention in the NLP community.[44] Nevertheless, automatic methods are able to achieve quite good performance, distinguishing between varieties like British and American English or Bosnian, Croatian, and Serbian with over 90% accuracy (Zampieri et al. 2014). When it comes to identifying Arabic varieties, some attention has been given to dialect identification, where automatic methods achieve somewhat lower accuracies of around 85-90% (Cotterell and Callison-Burch 2014; Zaidan and Callison-Burch 2014).

The most common technique in language identification, also adopted for dialect identification (Zaidan and Callison-Burch 2014), relies on language modeling. A *statistical* model of language assigns a probability to a text in the language. As a simple example, consider a language model that treats every word in the text independently from other words. In such a model, the probability of seeing a given word *w* is the number of occurrences of word *w* in the corpus, divided by the total number of words in the corpus:

$$P(w) = c(w)/C$$

Here $c(w)$ stands for the number of occurrences of word *w* in the entire corpus and C is the total number of words in the corpus. In such a model, the probability of a text with *K* words $w_1, w_2, \ldots w_K$ is the product of all the word probabilities:

$$P(w_1, w_2, \ldots w_k) = \prod_{k=1}^{K} P(w_k)$$

Note that such a model does not take the order or context of the words into account. It is based on individual word probabilities and is

[44] See Zampieri et al. 2014, and references therein.

therefore called a unigram language model. A higher-level bigram model looks at pairs of words and estimates the probability of a word given its previous words: $P(w_k|w_{k-1})$. In general, one might look at arbitrary word sequences of length N and construct an *N-gram* language model. In practice, there are all sorts of computational problems that need to be dealt with, such as assigning probability to low-frequency words, which is often solved by smoothing the probability distribution. Jurafsky and Martin (2008, Chapter 4) provide an overview of *N-gram* language models and associated techniques.

How could language models help in language or variety identification? Consider the following setting. Suppose we have several large corpora in known varieties, for instance, an MSA corpus, a Classical Arabic corpus, and a dialectal Arabic corpus. We can build a language model for each of these corpora, which are called the training corpora. Then, we are given a new text in an unknown variety, which is called the test text. We can use the language models to predict the probability that the text belongs to each of the three varieties. If the probability assigned to the new text by the MSA model is significantly higher than the probability assigned to it by the Classical or dialectal model, we decide that this is an MSA text. Now, we must take care to keep the training and test data distinct. If we have test data with known language varieties, we can measure how well our language models identify new texts. In practice, a common measure to assess the performance of language models is their *perplexity* on the test corpus. For a test text with words $w_1, w_2, \ldots, w_K$ the perplexity is defined based on the *N-gram* probability, normalized by the total number of words C:

The lower the perplexity, the higher the probability and the better the language model reflects the given text.

One last step is still missing: moving from individual variety identification to mixed text identification. In the present work, the approach explored builds on the assumption that a mixed text has characteristics of two or more language varieties. If these varieties are

more or less similarly represented in the mixed text,[45] then models constructed for individual varieties should be roughly equally successful in identifying them. In other words, models of two varieties represented in a mixed text should have similar perplexity values. To fully evaluate this proposed approach, a comprehensive evaluation on a corpus of mixed texts is needed, which is beyond the scope of the present work. Instead, a short experimental investigation is described below, along with some results.[46]

The first step in constructing the mixed texts identification experiment is to define training corpora for relevant language varieties. The following corpora, all described in Section 2, have been utilized. The chosen dataset for dialectal Arabic is mostly based on the human annotated dialectal parts of the Arabic Online Commentary (AOC) corpus by Zaidan and Callison-Burch (2011). The AOC includes sub-corpora for Egyptian, Jordanian, and Saudi dialects, collected from online user comments to news websites. In addition, the Egyptian Arabic Treebank (Maamouri et al. 2014), parts 1-3, was added to the Egyptian sub-corpus. The result is a corpus with high dialectal content, although it still has mixed elements.[47] For MSA, the Corpus of Contemporary Arabic (CCA) has been selected (Al-Sulaiti 2004); it contains mostly (though not solely) MSA texts in various domains. In addition, MSA texts from the AOC have been included in order to control for domain variation.[48] Finally, the King Saud University

[45] This is a rather strong assumption, as mixed texts can be situated anywhere on the continuum between dialectal and standard Arabic. While this assumption limits the current approach, it can nevertheless be applied to a large selection of mixed texts, and has empirical advantages.

[46] Preliminary results of this effort have been briefly mentioned in Arts et al. 2014, p. 369.

[47] For example, the Egyptian sub-corpus contains both standard Arabic negation *lam* and dialectal negation *ma-š*. Interestingly, their frequency is roughly similar (390 *lam*vs 361 *ma-š*), indicating a more dialectal language than certain contemporary spoken Mixed Arabic texts, where *lam* is much more frequent (Mejdell 2012, p. 241).

[48] Language identification is often sensitive to domain mismatch; a language model might learn to identify a difference in domain instead of language. Including texts from

Corpus of Classical Arabic (KSUCCA) was chosen for Classical Arabic (Alrabiah 2013). The sizes of the resulting corpora are given in Table 1.

Table 1: *Statistics of variety identification corpora*

	MSA	Classical	Egyptian	Jordanian	Saudi	Dialectal
Words	1.1M	48.7M	505K	286K	343K	1.1M
Letters	4.3M	201M	1.7M	988K	1.2M	3.9M

Next comes the construction of the language models. All of the experiments were conducted with the SRILM toolkit (Stolcke et al. 2002; Stolcke 2011).[49] After some initial experiments, a particular configuration seemed to be the most successful. This includes a 5-gram, character-level language model, that is, a language model that is trained on sequencers of letters rather than words.[50] A character-level language model operates on a smaller vocabulary (just the letters instead of every word), and a suitable smoothing technique is Witten-Bell discounting. All of these options are simple enough to specify with SRILM. Also note that the original Arabic text was pre-processed to separate punctuation and transliterate the Arabic to Latin script, which is more machine-readable.[51]

In order to make sure that the resulting language models are good, a small test set has been constructed, consisting of 30 texts in three varieties: MSA, classical, and Egyptian Arabic. The texts in this test set were selected to be relatively unmixed such that they reflect the individual varieties, to the extent that this is possible. Then, the models for all three varieties were applied on all the texts and the perplexity values were calculated. For each text, the model that had the lowest

similar domains in corpora used for different language models should alleviate this problem.

[49] http://www.speech.sri.com/projects/srilm/.

[50] Character-level language models are known to be quite successful in language identification (Zaidan and Callison-Burch 2014).

[51] Pre-processing was done with the MADA tool (Habash and Rambow 2005; Habash et al. 2009). Crucially, no tokenization or morphological analysis was applied, only transliteration and punctuation separation.

perplexity was chosen as the predicted variety. This procedure resulted in 93% accuracy, indicating that the built models are of reasonable quality.

Finally, let us consider how the built models can be used for finding mixed texts. The corpus selected for this experiment is arTenTen (Arts et al. 2014), a 5.8-billion-words corpus crawled from the web. Being comprised of web data, it contains texts in various genres and different varieties, so it is a reasonable candidate for our purposes. A portion of this corpus was extracted, containing about 3.4 million words in 4,807 texts. The language models for MSA, Classical, and dialectal Arabic were run on each of the texts. The results of the perplexity scores show that the vast majority of the texts, about 85%, were identified as MSA, with the rest split more or less equally between dialectal and Classical Arabic. Which of these texts are more likely to be mixed? One approach would be to look for texts whose perplexity scores for two varieties are very close to one another, but distant from the third variety. This approach had some success. For example, the following text had perplexity scores of 21.15, 22.15, and 46.2 for dialectal, MSA, and Classical Arabic, respectively. The text is mostly MSA, with a few non-standard elements: إنه with an initial *i* (can be interpreted as either *innahu* or *innu*, "that");[52] تفضلو with a missing final *alif* (*tafaḍḍalu* or *tfaḍḍalu*, "please, go ahead"); رايكم without a middle *hamza* (*raʾy* or *rāy*, "opinion"); مشكووورين ("thank you") with an elongation.[53]

نجم ستار أكاديمي 1 الكويتي بشار الشطي يعرض له حاليا من يومين
تقريبا كليب جديد رائع من حيث التصوير والإخراج والفكرة، الإخراج
للمخرج عزيز الجاسم شاب كويتي طموح أعتقد إنه عمله الأول، عموما

[52] All the transcriptions are approximate as inferring the correct pronunciation from the written text is not always possible. In fact, this is a known feature of mixed texts (e.g. Rosenbaum 2000, p. 72, 79).

[53] Most, if not all of the features mentioned in this section are well known in MMA texts, so they are not discussed in much detail here. For example, on the orthography of *alif* and *hamza* see Lentin 2008, p. 220; den Heijer 2012b, p. 161, and references therein.

> الكليب يحمل مفاجئة جميلة. لكل من لم يرى الكليب تفضلو على هذا الرابط وأعطوني رايكم. مشكوووورين…
>
> Star Academy [season] 1 star, the Kuwaiti Bashar Al-Shati, a new clip of his is now being shown since a couple of days ago, which has fabulous photography, production, and thought. The production is by the producer Aziz Al-Jasim, an ambitious Kuwaiti guy who I think it is his first work. Generally, the clip has a lovely surprise. For anyone who did not see the clip, go to this link and let me know what you think. Thaaaaank you.

As another example, consider the following excerpt from a text whose scores for dialectal and MSA were much better than its Classical Arabic score. The first sentence is in MSA while the last one has several dialectal elements, apparently Sudanese from the content: رايك without a middle *hamza* (as above); شنو (*šinu* or *šunu*, "what"), a common dialectal interrogative; تتباع (*titbāʿ*, "it is sold") with a passive inflection known, for example, from Egyptian Arabic; كلو (*kullu*, "all").

> وطبعا الجمهور عامة الذي ساند الفريق من داخل الاستاد فلقد أشدنا بهم من قبل في بوست منفرد. والحمدلله هلالاااااااااااااااااااااااااااااااااااااالاب. لكن ما قلت لى رايك شنو فى الدموع مع الكحل وجوطة[54] الكريمات لوحة سيريالية نادرة ممكن ما تتباع كلو كلو
>
> Of course, the general public that supported the team inside the stadium, we have already praised them before in a separate post. Thank god to Hilaalaaaaalaab.[55] But you did not tell me what you think of tears with kohl and the mix of creams, a rare surrealist painting that may not be sold at all.

However, when looking at more texts with two low perplexity values (usually for MSA and dialectal Arabic) compared to one higher score (usually Classical Arabic), it appears that many of them are not

---

[54] According to a Sudanese speaker, جوطة means "mix" in this context.

[55] *hilālāb* is a Sudanese form referring to the local football team Hilal. For example: https://hawamer.com/vb/hawamer622932-6 and http://www.skyscrapercity.com/showthread.php?p=53535399.

mixed at all. Rather, they are predominantly MSA. In fact, a more successful approach could be to simply look for texts identified as dialectal Arabic, that is, ones for which the perplexity score of the dialectal Arabic model was the lowest. Indeed, it can be argued that any *written* dialectal text is a mixed text to some extent. As an example, consider the following text, with highly dialectal content in Egyptian Arabic along with MMA norms: اوى, اووووووووى (*ʾawi*, "very"); كدا (*kida*, "this way"); نبقا with a final *alif* (*nibqa* or *nibʾa*, "we are"); يبداء with a final *hamza* (*yibdaʾ*, "he begins"); and so on.

> ياااااااااااااااااااااااه كلمات مؤلمة وحزينة اووووووووى الوداع بس ازاى وقت الوداع نهرب ونسيب الا بنودعهم اى نوع من الوداع هذا واى حب يهون اوى كدا وداع الموت اكيد لازم نبقا فى حضن الا بنودعة[56] وداع السفر اكيد برضو فى حضن الا مسافر وداع الحبايب اكيد بنستنى يمكن حد يبداء بكلمة ويتذكر كل واحد بحب الاخر وينسى الاساءه لعل وعسى يعود الحب فى القلوب من جديد
>
> My! These are painful and sad words, Dear me! Good bye only? How could we run away when the time of farewell comes without saying any good bye to them? Oh, what a poor love! It would be a deathly farewell! Of course, we must be close to each other and wish him a good trip, yes, be close and give the traveler the farewell of his friends. Yes, we will wait and may be someone will say the first word and everyone will remember he loves the other and will forget the evil he did. And perhaps may love fill the hearts again.

To conclude this section, it is important to acknowledge several limitations of the above approach. First, the automatic process is not perfect and identification mistakes happen. Second, the entire identification approach is done at the text level, and cannot identify the variety of specific words, phrases, and sentences. Some effort in identifying such sub-text units has been carried out in a number of recent studies (Solorio et al. 2014). For example, Elfardy et al. (2014) have combined language models with automatic morphological analysis and gazetteer lists, to perform word-level identification in

[56] To be read بنودعه.

code-switched data. It will be useful to combine such an approach with the one described here, perhaps by using their system as a second step of identifying language varieties in candidate mixed texts. Nevertheless, even with these caveats in mind, one could hope that automatic methods for variety identification will help MMA researchers search through large volumes of texts more rapidly in order to find and focus on specific interesting examples.

4. *Conclusion*

As more and more large-scale electronic Arabic corpora are becoming available, opportunities arise for new MMA research venues. Especially with the rise of electronic media, mixed language forms are becoming increasingly common. Computational methods can help researchers in studying these large collections of texts, whether by automatically identifying mixed texts or by providing large scale quantitative analyses. To this end, several important endeavors need to be tackled. A first important step may be in compiling electronic corpora that are dedicated to MMA, as most available corpora target other language varieties even if they do contain a significant number of MMA attestations. Such an effort will be especially fruitful if accompanied by the construction of linguistic databases of MMA features, as advocated in previous AIMA conferences. Another essential task is to develop natural language processing tools for MMA. Tools such as automatic morphological or syntactic analysis could spur quantitative studies on deeper linguistic levels. Such tools are largely lacking today, with the exception of a lemmatization tool presented by Tuerlinckx (2004). However, similar tools that target Arabic dialects are available and may prove useful for MMA texts.[57]

In a sense, these ideas are all part of an overarching goal of bringing together researchers from the MMA community and the natural language processing community. Hopefully, the present paper makes a modest step in setting the ground for future such collaboration for advancing the study of Middle and Mixed Arabic.

---

[57] Habash (2010, p. 141) mentions a few such tools; more recent tools include MADAMIRA, a morphological analyzer and disambiguator with support for Egyptian Arabic (Pasha et al. 2014).

References

Aguadé, Jorge, 2008. "Morocco", Versteegh, Kees, Mushira Eid, Alaa Elgibali, Manfred Woidich and Andrzej Zaborski (eds.), *Encyclopedia of Arabic Language and Linguistics*, Vol. 3, Leiden & Boston, Brill, p. 287-297.

Alansary, Sameh, Magdy Nagi and Noha Adly, 2007. "Building an International Corpus of Arabic (ICA): Progress of Compilation Stage", *Proceedings of the 7th International Conference on Language Engineering*, Cairo, Egypt, 30 p. https://www.researchgate.net/publication/242742358_Building_an_International_Corpus_of_Arabic_ICA_Progress_of_Compilation_Stage

Almeman, Khalid and Mark Lee, 2013. "Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words", *1st International Conference on Communications Signal Processing, and their Applications (ICCSPA)*. American University of Sharjah (Technical Sponsor IEEE), p. 1-6.

Alrabiah, Maha, AbdulMalik Al-Salman and Eric Atwell, 2013. "The Design and Construction of the 50 Million Words KSUCCA", *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, Lancaster University, UK, The University of Leeds, p. 5-8. https://pdfs.semanticscholar.org/544b/b65282fb852f1ba9b960-1c95d547d38659ed.pdf

Al-Sulaiti, Latifa, 2004. *Designing and Developing a Corpus of Contemporary Arabic*, M.Sc. Thesis, The University of Leeds, Leeds, UK.

Al-Thubaity, Abdulmohsen O., 2015. "A 700M+ Arabic corpus: KACST Arabic corpus design and construction", *Language Resources and Evaluation*, Vol. 49, No. 3, p. 721-751.

Arts, Tressy, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff and Vit Suchomel, 2014. "arTenTen: Arabic Corpus and Word Sketches", *Journal of King Saud University - Computer and Information Sciences*, Vol. 26, No. 4, p. 357-371.

Belinkov, Yonatan, Alexander Magidow, Maxim Romanov, Avi Shmidman and Moshe Koppel, 2016. "Shamela: A Large-Scale

Historical Arabic Corpus", *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, Osaka, Japan, p. 45-53.

Bouamor, Houda, Nizar Habash and Kemal Oflazer, 2014. "A Multidialectal Parallel Corpus of Arabic", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, European Language Resources Association (ELRA), p. 1240-1245.

Cotterell, Ryan and Chris Callison-Burch, 2014. "A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, European Language Resources Association (ELRA), p. 241-245.

den Heijer, Johannes, 2012a. "Introduction: Middle Arabic and Mixed Arabic, A New Trend in Arabic Studies", Zack, Liesbeth and Arie Schippers (eds.), *Middle Arabic and Mixed Arabic: Diachrony and Synchrony*, Leiden & Boston, Brill, p. 1-25.

den Heijer, Johannes, 2012b. "Towards an Inventory of Middle and Mixed Arabic Features: The Inscriptions of Deir Mar Musa (Syria) as a Case Study", Zack, Liesbeth and Arie Schippers (eds.), *Middle Arabic and Mixed Arabic: Diachrony and Synchrony*, Leiden & Boston, Brill, p. 157-173.

Elewa, Abdel-Hamid, 2004. *Collocation and Synonymy in Classical Arabic: A Corpus-based Approach*, Ph.D. Thesis, The University of Manchester, Manchester, UK.

Elfardy, Heba, Mohamed Al-Badrashiny and Mona Diab, 2014. "AIDA: Identifying Code Switching in Informal Arabic Text", *Proceedings of the First Workshop on Computational Approaches to Code Switching*, Doha, Qatar, Association for Computational Linguistics, p. 94-101.

Habash, Nizar Y., 2010. *Introduction to Arabic Natural Language Processing*, Synthesis Lectures on Human Language Technologies, San Rafael, Morgan & Claypool Publishers.

Habash, Nizar and Owen Rambow, 2005. "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop", *Proceedings of the Association for Computational Linguistics*, Ann Arbor, Michigan, p. 573-580.

Habash, Nizar, Owen Rambow and Ryan Roth, 2009. "MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization", *Proceedings of the International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt, p. 102-109.

Hary, Benjamin, 2003. "Judeo-Arabic: A Diachronic Reexamination", *International Journal of the Sociology of Language* 163, 61-75.

Hary, Benjamin, 2009. *Translating Religion: Linguistic Analysis of Judeo-Arabic Sacred Texts from Egypt*, Leiden & Boston, Brill.

Hary, Benjamin, 2012. "Judeo-Arabic as a Mixed Language", Zack, Liesbeth and Arie Schippers (eds.), *Middle Arabic and Mixed Arabic: Diachrony and Synchrony*, Leiden & Boston, Brill, p. 125-143.

Daniel Jurafsky and James H. Martin, 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* [2nd ed.], Upper Saddle River, New Jersey, Pearson - Prentice Hall. https://www.cs.colorado.edu/~martin/SLP/Updates/1.pdf

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell, 2004. "The Sketch Engine", *Proceedings of EURALEX*, Lorient, France, p. 105-116. http://www.euralex.org/elx_proceedings/Euralex2004/011_-2004_V1_Adam%20KILGARRIFF,%20Pavel%20RYCHLY,%20Pavel%20SMRZ,%20David%20TUGWELL_The%20%20Sketch%20Engine.pdf

Lentin, Jérôme, 2008. "Middle Arabic", Versteegh, Kees, Mushira Eid, Alaa Elgibali, Manfred Woidich and Andrzej Zaborski (eds.), *Encyclopedia of Arabic Language and Linguistics*, Vol. 3, Leiden & Boston, Brill, p. 215-224.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter and Wigdan Mekki, 2004. "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus", *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 8 p.

https://www.researchgate.net/publication/228693973_The_-penn_arabic_treebank_Building_a_large-scale_annotated_arabic_-corpus

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow and Dalila Tabessi, 2006. "Developing and Using a Pilot Dialectal Arabic Treebank", *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genova, Italy, European Language Resources Association (ELRA), p. 443-448.

Maamouri, Mohamed, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash and Ramy Eskander, 2014. "Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, European Language Resources Association (ELRA), p. 2348-2354.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz, 1993. "Building a Large Annotated Corpus of English: The Penn Treebank". *Computational Linguistics* 19(2), p. 313-330.

Meftouh, Karima, Salima Harrat, Salma Jamoussi, Mourad Abbas and Kamel Smaili, 2015. "Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus", *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, p. 26-34.
http://www.aclweb.org/anthology/Y15-1004

Mejdell, Gunvor, 2012. "Playing the Same Game? Notes on Comparing Spoken Contemporary Mixed Arabic and (Pre)Modern Written Middle Arabic", Zack, Liesbeth and Arie Schippers (eds.), *Middle Arabic and Mixed Arabic: Diachrony and Synchrony*, Leiden & Boston, Brill, p. 235-245.

Mubarak, Hamdy and Kareem Darwish, 2014. "Using Twitter to Collect a Multi-Dialectal Corpus of Arabic", *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing*, Doha, Qatar, Association for Computational Linguistics, p. 1-7.

Parker, Robert, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda, 2011. *Arabic Gigaword Fifth Edition LDC2011T11*, Web Download, Philadelphia, Linguistic Data Consortium.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth, 2014. "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, European Language Resources Association (ELRA), p. 1094-1101. http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf

Rashwan, Mohsen A.A., Mohamed A.S.A.A. Al-Badrashiny, Mohamed Attia, Sherif M. Abdou and Ahmed Rafea, 2011. "A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Unfactorized Textual Features", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, no. 1, p. 166-175.

Romanov, Maxim G., 2013. *Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunni World (661-1300 CE)*, Ph.D. Dissertation, University of Michigan, Michigan.
https://deepblue.lib.umich.edu/handle/2027.42/102300

Rosenbaum, Gabriel M., 2000. "*'Fuṣḥāmmiyya'*: Alternating Style in Egyptian Prose", *Zeitschrift für Arabische Linguistik* 38, p. 68-87.

Saad, Motaz K. and Wesam Ashour, 2010. "OSAC: Open Source Arabic Corpora", *6th International Conference on Electrical and Computer Systems*, Cyprus, p. 118-123.

Sadat, Fatiha, Farzindar Kazemi and Atefeh Farzindar, 2014. "Automatic Identification of Arabic Language Varieties and Dialects in Social Media", *Proceedings of the Second Workshop on Natural Language Processing for Social Media* (SocialNLP), Dublin, Ireland, Association for Computational Linguistics and Dublin City University, p. 22-27.

Salama, Ahmed, Houda Bouamor, Behrang Mohit and Kemal Oflazer, 2014. "YouDACC: the Youtube Dialectal Arabic Comment Corpus", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, European Language Resources Association (ELRA), p. 1246-1251.

Shoufan, Abdulhadi and Sumaya Al-Ameri, 2015. "Natural Language Processing for Dialectal Arabic: A Survey", *Proceedings of*

the Second Workshop on Arabic Natural Language Processing, Beijing, China, Association for Computational Linguistics, p. 36-48.

Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang and Pascale Fung, 2014. "Overview for the First Shared Task on Language Identification in Code-Switched Data", *Proceedings of the First Workshop on Computational Approaches to Code Switching*, Doha, Qatar, Association for Computational Linguistics, p. 62-72.
https://pdfs.semanticscholar.org/9d14/70698bbf4573974c-97b62ad466fe3127f7fd.pdf

Stolcke, Andreas, 2002. "SRILM – An Extensible Language Modeling Toolkit", *Proceedings of the International Conference on Spoken Language Processing*, Vol. 2, Denver, Colorado, p. 901-904.

Stolcke, Andreas, Jing Zheng, Wen Wang and Victor Abrash, 2011. "SRILM at Sixteen: Update and Outlook", *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawai, 5 p.
https://www.researchgate.net/publication/255563494_-SRILM_at_sixteen_update_and_outlook

Tuerlinckx, Laurence, 2004. "La lemmatisation de l'arabe non classique", *Le poids des mots. Actes des JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles*, Vol. 2, p. 1069-1078.
http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_105.pdf.

Versteegh, Kees, 2006. *The Arabic Language*, Edinburgh, Edinburgh University Press.

Zaghouani, Wajdi, 2014. "Critical Survey of the Freely Available Arabic Corpora", *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tool*, Reykjavik, Iceland, LREC, p. 1-8.

Zaidan, Omar and Chris Callison-Burch, 2011. "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers*, Portland, Oregon, Association for Computational Linguistics, p. 37-41.

Zaidan, Omar and Chris Callison-Burch, 2014. "Arabic Dialect Identification", *Computational Linguistics*, Vol. 40, No. 1, p. 171-202.

Zampieri, Marcos, Liling Tan, Nikola Ljubešić and Jörg Tiedemann, 2014. "A Report on the DSL Shared Task 2014", *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties, and Dialects*, Dublin, Ireland, Association for Computational Linguistics and Dublin City University, p. 58-67. http://www.aclweb.org/anthology/W14-5307

Zemanek, Petr, 2001. "CLARA (Corpus Linguae Arabicae): An Overview", *Proceedings of ACL/EACL Workshop on Arabic Language Processing: Status and Prospects*, p. 111-112.