Tharwa: A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Pradeep Dasigi, Heba Elfardy[†], Ramy Eskander[†], Nizar Habash[†], Abdelati Hawwari, Wael Salloum[†]

Department of Computer Science, The George Washington University, Washington, DC

[†]Center for Computational Learning Systems, Columbia University, New York, NY

mtdiab@gwu.edu

Abstract

We introduce an electronic three-way lexicon, Tharwa, comprising Dialectal Arabic, Modern Standard Arabic and English correspondents. The paper focuses on Egyptian Arabic as the first pilot dialect for the resource, with plans to expand to other dialects of Arabic in later phases of the project. We describe Tharwa's creation process and report on its current status. The lexical entries are augmented with various elements of linguistic information such as POS, gender, rationality, number, and root and pattern information. The lexicon is based on a compilation of information from both monolingual and bilingual existing resources such as paper dictionaries and electronic, corpus-based dictionaries. Multiple levels of quality checks are performed on the output of each step in the creation process. The importance of this lexicon lies in the fact that it is the first resource of its kind bridging multiple variants of Arabic with English. Furthermore, it is a wide coverage lexical resource containing over 73,000 Egyptian entries. Tharwa is publicly available. We believe it will have a significant impact on both Theoretical Linguistics as well as Computational Linguistics research.

Keywords: Egyptian Arabic Dictionary, Arabic Dialects, Arabic Morphology, Arabic Lexicon

1. Introduction

The Arabic language is an aggregate of multiple varieties including a standard used in education and official settings known as Modern Standard Arabic (MSA) and a number of spoken vernaculars comprising the dialectal variants of the language, collectively known as Dialectal Arabic (DA). DA are emerging as a significant set of language varieties for textual processing due to their pervasive and ubiquitous presence online especially in the current influx of social media. The differences between DAs and MSA go beyond register differences as is typical in other languages (formal vs. informal). Coarsely, the two varieties of Arabic, MSA and DA, co-exist in a state of diglossia (Ferguson, 1959), in a relative complementary distribution but crucially they differ significantly from one another on the morphological, phonological and lexical levels of linguistic representation. Such differences have a direct impact on Arabic processing tools. Most automatic resources exist for MSA leading to an abundance of tools for processing this variety but given the significant difference between MSA and DA, we note a sharp drop in performance for the tools when applied to DA. Differences on the lexical level are especially interesting since many surface word forms are homographically similar across naturally occurring written Arabic variants in particular in the absence of short vowel representation –aka diacritics. Many of these forms are not semantic cognates which leads to significant deterioration in computational performance. To date, a notable gap exists for DA resources especially ones that bridge across variants and to English. Some computational approaches to dialectal processing such as Abo Bakr et al. (2008) and Salloum and Habash (2011) have addressed the gap by approximations via extending BAMA/SAMA databases (Buckwalter, 2004; Graff et al., 2009) to accept DA prefixes and suffixes. This is, however, a shallow process that is limited to a subset of the lexicon shared by both MSA and DA. Hence, the creation of different resources such as lexicons is crucial from a computational point of view. Moreover, linguistically, a resource that fills this lexical gap can lead to more thorough analysis of DA content leading to better insights into the nature of these varieties and how they are being used and what is their exact relation with MSA. This could potentially lead to interesting research in theoretical linguistics, sociolinguistics, comparative linguistics, lexical semantics, lexicography and discourse analysis.

Here we introduce Tharwa, a three-way lexicon between the Egyptian variety of DA, Egyptian Arabic (EGY), MSA and English (ENG). In addition to providing word level equivalents across the Arabic varieties and their correspondences in ENG, Tharwa provides rich linguistic information for each entry such as part of speech (POS), number, gender, rationality, and morphological root and pattern forms. Tharwa is primarily a lemma based resource, namely all the DA and MSA and ENG entries are chosen conventionalized citation forms. Tharwa is the first resource of its kind for Arabic. It currently serves as a nucleus to be extended to other Arabic variants. Tharwa is based on a compilation of information from both monolingual and bilingual existing resources such as paper dictionaries and electronic, corpus-based dictionaries. Multiple levels of quality checks are performed on the output of each step in the creation process. The importance of Tharwa lies in the fact that it is the first resource of its kind bridging multiple variants of Arabic with English. Furthermore, it is a wide coverage lexical resource containing over 73,000 EGY entries.

2. Egyptian Arabic

In characterizing DA, EGY stands in a cluster of its own due to its significant difference from MSA and other Arabic varieties (Brustad, 2000). It is one of the most widespread varieties of Arabic due to the fact that it is the native tongue of more than 90 million contemporary Arabs (which makes up for close to one third of the Arabic-speaking world), along with the strategic and cultural importance of Egypt, but also the media impact of Egypt is quite widespread leading to EGY being very well understood by most non-Egyptian Arabs. EGY exhibits considerable differences from MSA at multiple levels of linguistic representation. We will briefly address here only the morphological, phonological, and lexical variation from MSA without touching upon the syntactic differences. For more information on EGY differences from MSA, see (Habash et al., 2012b).

2.1. Phonological Variation

As is the case for many languages and their dialects, the pronunciation of some MSA phonemes have shifted in EGY. Some of the shifts are quite regular such as /q/ (of the letter ق) becoming a glottal stop /'/ except for few words borrowed from MSA or Classical Arabic, e.g., the word قلب 'heart' is pronounced /qalb/ in MSA but /'alb/ in EGY. Another example is the MSA /θ/ phoneme (of the letter ث) which shifts in some words to /t/ and in others to /s/, e.g., ثلاثة *vlAvp* 'three' is pronounced as MSA /θala:θa/ or EGY /tala:ta/, and ثروة *vrwp* 'wealth, fortune' is pronounced as MSA /θarwa/ and EGY /sarwa/. The differences in the phonology affect how people write, especially given the absence of an orthographic standard for EGY. In our work, we use the conventional orthography for dialectal Arabic (CODA) proposed for EGY by Habash et al. (2012b), but we recognize common alternative spellings also.

2.2. Morphological Variation

EGY morphology exhibits considerable divergence from MSA in both inflectional and derivational morphology. We note that the derivational differences are more relevant for building a lexical resource such as Tharwa; however we will review some of the inflectional variations. For an extensive discursive of Arabic morphology in NLP, see (Habash, 2010).

Affixation EGY has some unique prefixes, suffixes and clitic morphemes that are not shared by MSA, e.g., the EGY future tense prefixes +ه *ha*+[1] and +ح *Ha*+ are notably different from the MSA future prefix +س *sa*+.

Case inflection While MSA has a complex case system, EGY does not. Different inflected forms in MSA map to the same form in EGY, e.g., MSA موظفون *mwZfwn*, 'employees [nom.]' and MSA موظفين *mwZfyn*, 'employees [acc./gen.]' map to EGY موظفين *mwZfyn*, 'employees'.

Derivational differences MSA and EGY have similar word formation mechanisms, particularly because derivational morphology depends on roots and patterns. However, EGY has some morphological patterns which are not used in MSA such as `AisotaC1aC2C2aC3`, e.g. استخبّى *Aisotaxab∼aY* 'to hide'. In addition EGY utilizes non-MSA morphological patterns to represent the passive voice or the unaccusative form of some verbs such as `AitoC1aC2aC3` (e.g. اتكتب *Aitokatab* 'to be written'), `AitoC1aC2C2aC3` (e.g. اتصوّر *AitoSaw∼ar* 'to have his picture taken'), and `AitoC1AC2iC3` (e.g. اتّاكل *Ait∼Akil* 'to be eaten').

2.3. Lexical Variation

The EGY lexicon comprises entries that differ as well as overlap with MSA:

Identical EGY and MSA words that are identical in all respects phonological, orthographic, morphological, and semantic, e.g. نشيط *na$iyT*, 'active'.

Semantic Cognates EGY and MSA that share the same meaning but with some regular phonological and/or orthographic variation, e.g., EGY verb لعب *liEib* 'to play' corresponds to MSA verb لعب *laEib*.

Homographs/Homophones EGY and MSA that have the same orthography and pronunciation but different meanings, e.g. حاجة *HAjap* is 'necessity' in MSA, but could mean both 'thing' as well as 'necessity' in EGY.

Distinct Words that belong uniquely to only one of the varieties EGY or MSA, e.g. مش *mi$* 'not', بس *bas* 'only, enough', and دغري *dugoriy* 'straight-ahead' are only used in EGY.

3. Related Work

Unlike MSA, EGY has a small number of printed (bilingual or monolingual) Dictionaries. (Spiro, 1895) is the first recorded dictionary of the Egyptian dialect, and its modern reproduction (Spiro, 1987) contains 12,500 EGY-ENG entries. (1986) compiled *A Dictionary of Egyptian Arabic* which is the most comprehensive and complete dictionary in print for EGY, consisting of more than 31K EGY-ENG single word entries. Both dictionaries target English non-Arabic speakers learning EGY.

Machine Readable Dictionaries (MRDs) of EGY have appeared with varying degrees of coverage and linguistic sophistication. The Egyptian Colloquial Arabic Lexicon (ECAL) by the Linguistic Data Consortium (LDC) (Kilany et al., 2002) is a monolingual lexicon of fully inflected words (surface forms) that consists of over 66K monolingual entries. ECAL is used by (Habash et al., 2012a) to produce the CALIMA morphological analyzer for EGY. The Columbia Egyptian Colloquial Arabic Dictionary (CECAD) (Maamouri et al., 2006) is a small EGY-MSA-ENG dictionary consisting of 1,752 high frequency words. CECAD is a subset of ECAL manually augmented with MSA and ENG equivalents.

4. Building Tharwa

Tharwa is a three-way EGY-MSA-ENG lemma based lexicon augmented with morphosyntactic and morphosemantic information pertaining to the EGY entry. A lemma is defined as a 3rd person singular masculine form for nominals and the perfective 3rd person form for verbal entries. The additional linguistic information is limited to part of speech (POS), gender, number, rationality, morphological pattern, and morphological root. The resource also includes closed class words and some named entities. Since there exist no standard orthography for DA in general, we use a standardized written form for EGY based on CODA (Habash et al., 2012b) as the pivot for an entry in Tharwa, however

[1] Arabic transliteration is in the Buckwalter scheme (Habash et al., 2007).

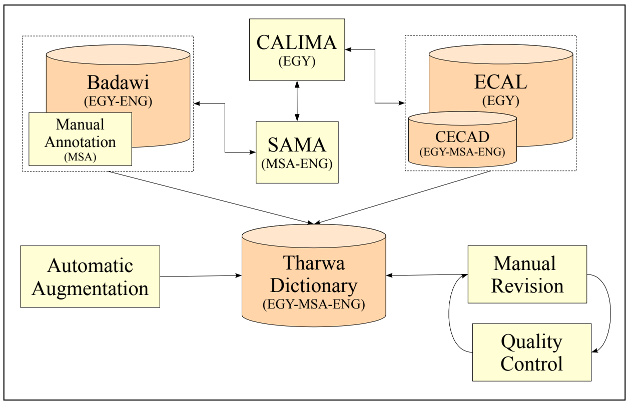


Figure 1: Tharwa Building Process

we include as many orthographic variants for the EGY entry as possible. Moreover, the EGY and MSA entries are fully diacritized to reflect the phonology and morphology explicitly. The creation of Tharwa relied on several preexisting heterogeneous resources which were merged and the gaps filled to achieve the desired information profile for each entry in the lexicon. The gaps are filled manually and semi-automatically. The resource is subjected to an iterative quality control process for consistency both manually and automatically.

4.1. Pre-existing Lexical Resources

Hinds-Badawi Dictionary (BADAWI) The Hinds-Badawi Dictionary (Badawi and Hinds, 1986) (BADAWI) is a paper-based EGY-ENG dictionary that provides EGY word entries with their corresponding English translations and definitions. The EGY entries are written in both undiacritized Arabic script and a full phonological transcription close to IPA. Each EGY entry is associated with a coarse grained POS category such as noun or verb. The dictionary comprises 31,548 EGY-ENG single word entries.[2]

Egyptian Colloquial Arabic Lexicon (ECAL) ECAL is a machine readable monolingual lexicon (Kilany et al., 2002). It was developed by the LDC as part of the CALLHOME Egyptian Arabic (CHE) corpus (Gadalla et al., 1997). ECAL has ∼66K EGY entries, each of which consists of a phonological form, an undiacritized Arabic script orthography form, a lemma (in phonological form), and morphological features. ECAL has ∼36K unique lemmas.

Columbia Egyptian Colloquial Arabic Dictionary (CECAD) CECAD is a three-way, EGY-MSA-ENG small lexicon developed at Columbia University as an extension to a portion of ECAL (Maamouri et al., 2006). CECAD consists of 1,752 entries extracted from the top most frequent entries in ECAL. The entries are manually augmented with MSA and ENG correspondents using entries from the BAMA morphological analyzer (Buckwalter, 2004), a predecessor of the SAMA analyzer (see below).

CALIMA Lexicon (CALIMA-LEX) CALIMA (aka CALIM-ARZ) is an EGY morphological analyzer (Habash et al., 2012a). At its core, CALIMA relies on the ECAL lexicon after undergoing several rounds of orthographic conventionalization, part-of-speech mapping, and morphological segmentation. All details of how ECAL is used to build CALIMA are discussed in Habash et al. (2012a). CALIMA has only EGY entries covering ∼36K unique lemmas, which is the full list of ECAL lemmas.

SAMA Lexicon (SAMA-LEX) The Standard Arabic Morphological Analyzer (SAMA) (Graff et al., 2009) is a morphological analyzer for MSA developed by the LDC. Its internal engine utilizes an lexical database that consists of ∼41K MSA lemmas along with their English equivalents and LDC standard POS tags. Some of the SAMA lemmas are disambiguated semantically using id markers, although not completely. We ignore these markers, which reduces the total number of lemmas to ∼37K.

4.2. Tharwa Compilation

Tharwa at its core is a merge of all the above resources; however, these resources exist in different formats with different partial information. We next discuss the different processes we apply to all the resources to standardize the forms of the EGY entries, provide them with MSA and ENG lexical correspondents, and augment them entries with morphosyntactic and lexico-semantic information.

Egyptian Entry Standardization All the resources go through a process of standardization for the EGY entries to be rendered CODA compliant. For Tharwa purposes, entries from the BADAWI paper dictionary are manually copied by bilingual native Arabic annotators with proven proficiency in EGY, MSA and ENG. All the annotators had at least BA degrees. The annotators are asked to render the English as equivalents, mark the EGY entry as lemma or surface form, provide equivalent MSA synonym(s). They were specifically instructed to ignore EGY MWE entries. They only copied the undiacritized form of the EGY entry (main entry, not the Romanized IPA variant). CALIMA-LEX entries are already CODA compliant as part of the CALIMA internal clean up process. We use only the ECAL set cleaned up and used within CALIMA-LEX. However, we maintain the original ECAL orthographic forms for the EGY entries in addition to the diacritized CODA form. In this standardization step, the entries are rendered fully diacritized. All the surface forms from BADAWI are kept but we add their lemmas as additional entries if absent. We also maintain the original form whether it was an inflected surface form or a lemma in the original orthography. We perform a round of clean up on the entries correcting for spelling mistakes especially paying particular attention for Hamza variants, and Alif-Maqsura ى *Y* versus Ya ي *y* cases. Providing the full diacritization for each of the entries allows for variation in the POS tag associated with an entry. The undiacritized forms pack several POS tags in addition to the semantic homonyms and synonyms. For instance, the entry أمر *>mr* as rendered in the original BADAWI lexicon once diacritized is split into three different entries: (a) adjective, *>amar* 'more bitter', (b) noun, *>amor* 'order', and (c) verb, *>amar* 'to order'. As mentioned above, some of the resources provide POS tag information. However this POS information relies on different tag sets. We adopt the uninflected POS tag set (different from the inflected Buckwalter tagset) in the SAMA-LEX

[2] We are not reporting on the multiword expression entries.

and CALIMA-LEX as used in MADAMIRA (Pasha et al., 2014), for MSA and EGY, respectively. We merge the resources providing the union of all the entries. All the Arabic entries (EGY and MSA) in the resources are rendered in both UTF8 Arabic script as well as Buckwalter transliteration encoding. We maintain the source information for each entry. This process results in a merged CODAfied fully diacritized pos tagged version of the resources comprising BADAWI, CECAD, and CALIMA-LEX.

MSA Augmentation As mentioned above, BADAWI Dictionary, ECAL and CALIMA-LEX do not provide MSA equivalents in their original form, hence we manually augment them with MSA equivalents. We provide the MSA correspondents for both EGY lemmas and the inflected surface forms resulting from the previous step of standardization. The equivalent for an EGY inflected surface form is an MSA inflected surface form or set of inflected surface forms and the correspondent for an EGY lemma is an MSA lemma or set of lemmas. We also manually provide the MSA POS tag information for the MSA equivalent(s). As mentioned earlier, we adopt the LDC standard MSA POS tag set provided with SAMA. Once the MSA equivalents are generated, the resulting combined lexicon with EGY and corresponding MSA and POS information goes through a process of exhaustive manual checking and then automatic compression to eliminate redundancies while maintaining source information. The unification process is not trivial. It is performed on two levels: coarse grained unification and a fine grained unification. The initial fine grained unification is performed on the fully diacritized form of the EGY entries coupled with their POS information creating a unique primary key. However, we note that the diacritization step is manually created which is prone to errors, hence we utilize a coarse unification step where the EGY entry is stripped of the diacritization while maintaining the POS information. The resulting merged entries are further subjected to manual checking to ensure that the merge of entries is warranted. SAMA-LEX provided MSA with ENG equivalents as well as POS tag information. We merge the resulting combined lexicon with SAMA-LEX pivoting on the MSA entries coupled with POS information. The set that is not matched from SAMA, i.e. where the MSA entry is uniquely in SAMA but does not exist in the combined lexicon, is augmented with EGY entries in lemmatized diacritized CODA form together with its corresponding POS tag information. Another round of merging takes place, then a round of manual quality control check is applied to the resulting resource, BCS which has EGY, MSA, and POS tags.

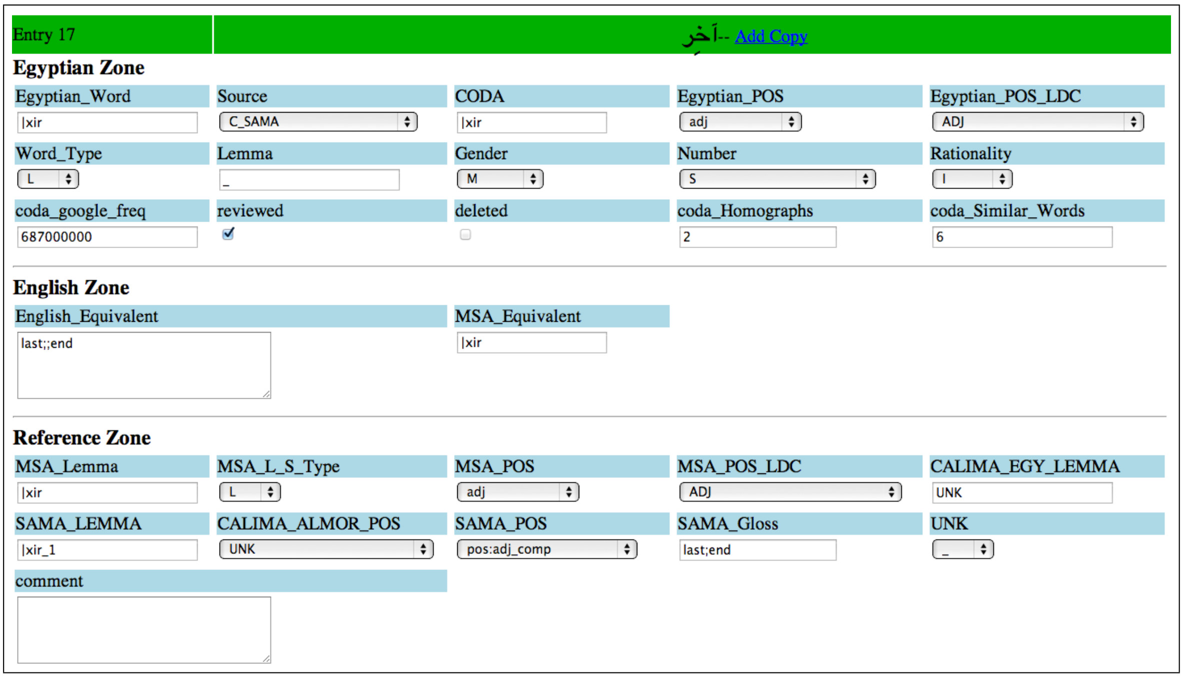
English Augmentation The original BADAWI, CECAD and SAMA provide English correspondents. However, they are not always exact equivalents, in some cases they are definitions, not in lemma form, or include homonyms. The merged BCS resource resulting from the previous two steps of standardization and MSA augmentation has a set of English correspondents in most of the entries from one of the sources, however, some entries do not have ENG equivalents. Hence we fill all the missing equivalents and perform several rounds of clean up on the ENG information by rendering equivalent correspondents (lemmas and inflected surface forms for EGY lemmas and EGY inflected surface forms, respectively). We split semantic homonym entries into multiple EGY entries. Unlike other relevant lexical resources such as BAMA (Buckwalter, 2004) and SAMA, we adopt a medium-grained approach to word senses, exploiting the theoretical distinction between synonyms, polysemes and homonyms. If the English equivalents describe the same sense they are considered synonyms and interchangeable equivalents of the same lexical entry. Otherwise if the English equivalents describe different senses (whether the senses are extensions of each other, i.e., polysemes, or not, i.e. homonyms) each distinct sense warrants a split creating separate EGY entries. To illustrate, the entry بيت *bayt* has two meanings ‘house’ and ‘verse’ suggesting two homonyms (homophones and homographs), hence it is split based on the ENG equivalents into two separate entries. Similarly, the entry عامل *EAmil* ‘worker’ and ‘factor’ suggests two polysemes and hence is split into two different entries. These examples are contrasted against حاير *HAyir*, ‘confused’, ‘uncertain’, and ‘baffled’, all of which are synonymous, hence, the English equivalents are not split out. The process is performed manually with multiple rounds of quality control checks for consistency in format. The bilingual annotators are specifically instructed to render the English equivalent(s) by inspecting the EGY and MSA and corresponding POS information, simultaneously.

Augmentation with Linguistic Information Extra linguistic information is added to each EGY entry including:

- Word type as in lemma or inflected surface form, e.g., compare the EGY noun lemma عين *Eayn* ‘eye’ to its broken (irregular) plural inflected form عيون *Euyuwn* ‘eyes’ (which is also linked to its lemma).
- The semantic attributes of gender, number, and rationality where applicable, e.g., the inflected entry عيون *Euyuwn* ‘eyes’ is marked as feminine, plural and irrational. We follow the conventions for marking these attributes proposed by Alkuhlani and Habash (2011).
- Morphological pattern and root information where applicable, e.g., the EGY verb lemma استبدل *Aisotabodil* ‘to change’ has the root *bdl* and pattern `AisotaC1oC2iC3`.

4.3. Manual Augmentation Process

Graphical User Interface We developed a web based application (Tharwa-GUI), as illustrated in Figure 2, designed specifically for the purpose of managing, maintaining, updating and extending Tharwa. Tharwa-GUI provides a Graphical User Interface for human lexicographers to review, modify and update lexical entries and their associated morpho-syntactic and semantic information. To ensure consistency while minimizing accidental data editing errors, Tharwa-GUI relies on controlled user input methods, such as check boxes and drop-down lists for most of the fields. However several of the fields in Tharwa are free form, hence some of the checks performed by the annotators involve: a. Converting MSA and ENG definitions and paraphrases into lemmas; b. Ensuring that all MSA equivalents of an EGY token have the same POS tag, same number and gender for nouns and adjectives, and same tense and voice for verbs; c. Ensuring that the EGY and MSA are correctly diacritized; d. Ensuring that the ENG is indeed the



Entry 17	آخِر-- Add Copy

Egyptian Zone

Egyptian_Word	Source	CODA	Egyptian_POS	Egyptian_POS_LDC
|xir	C_SAMA	|xir	adj	ADJ
Word_Type	**Lemma**	**Gender**	**Number**	**Rationality**
L	_	M	S	I
coda_google_freq	**reviewed**	**deleted**	**coda_Homographs**	**coda_Similar_Words**
687000000	☑	□	2	6

English Zone

English_Equivalent	MSA_Equivalent
last;;end	|xir

Reference Zone

MSA_Lemma	MSA_L_S_Type	MSA_POS	MSA_POS_LDC	CALIMA_EGY_LEMMA
|xir	L	adj	ADJ	UNK
SAMA_LEMMA	**CALIMA_ALMOR_POS**	**SAMA_POS**	**SAMA_Gloss**	**UNK**
|xir_1	UNK	pos:adj_comp	last;end	_
comment				

Figure 2: Tharwa-GUI: the Entry Editor Web Application

correct correspondent; e. Removing partial non-sensical multi-word expressions (MWE) singleton words that never occur on their own in EGY such as سداح *sadAH* which is part of the MWE سداح مداح *sadAH madAH* 'state of confusion/chaos'. Tharwa-GUI relies on a backend MySQL database. The backend database organized structure takes as primary key the combination of CODAfied diacritized EGY entry, ENG, MSA and POS creating a unique entry. The existence of the database allows for the ease of generating frequency counts and statistics pertaining to the different attributes as well as data subsets.

Crowd Sourcing We are currently harvesting MSA and ENG equivalents for existing EGY entries leveraging the power of crowd sourcing. We have designed experiments for both verification and augmentation. In the verification phase we have *rating* experiments where the annotators are asked to indicate whether a triple EGY-MSA-ENG is correct or not, i.e., a binary decision. We also have *generating* experiments where we provide the annotators with two of the fields and ask them to generate the third. Hence we present the crowd with the EGY and the MSA and ask them to provide the ENG, or the EGY and ENG and ask them to provide the MSA. It is worth noting that we provide the Arabic, EGY and MSA, fully diacritized. In the case of EGY, we exhaustively provide all the orthographic variants we have in Tharwa, not only the CODA form. We submit these variants as separate instances for the purposes of annotation. For example, in rating experiments the EGY lemma ثلاثة *valAvap* is presented to the annotators as the two (EGY, MSA, ENG) triples: (ثلاثة *valAvap*, ثلاثة *valAvap*, 'gold') (CODA) and (تلاتة *talAtap*, ثلاثة *valAvap*, 'gold') (non-CODA). The annotators are asked to rate if these are correctly corresponding to each other, i.e. the triples are correct. In the ENG generating experiments, we present the annotators with the following two pairs (EGY, MSA): (ثلاثة *valAvap*, ثلاثة *valAvap*) (CODA) and (تلاتة *talAtap*, ثلاثة *valAvap*) (non-CODA). The annotators are asked to provide the ENG correspondent(s). In the augmentation step, we provide the annotators with the MSA and ENG equivalents and ask them to provide the EGY equivalent.[3] We also created a variant of these crowd sourcing experiments where we present the annotators with the words in context with example sentences extracted from lemmatized diacritized parallel corpora utilized specifically as described in Section 4.4.. This verification and augmentation steps are still work in progress.

4.4. Automatic Verification and Augmentation via Parallel Corpora

We exploit parallel corpora that exist for EGY-ENG and MSA-ENG in the process of verifying and augmenting the manual process of Tharwa creation. We derive word level correspondents via automatic word-level alignment applied on lemmatized parallel corpora. This approach in principle is similar to that taken by Saleh and Habash (2009) for learning lemma-based dictionaries from parallel data, however we triangulate two parallel corpora simultaneously. We use *Bolt-ARZ* $v4 + v3$ (LDC2012E89 and LDC2012E99) for EGY-ENG parallel data. This data contains 3.5 million EGY words. For MSA-ENG parallel data, we use *GALE* $phase4$ (LDC2008E22) data which contains ~60 million MSA words. We focus on the automatic verification and augmentation of three main fields in Tharwa, namely: EGY CODA lemma, MSA lemma, and ENG equivalent. Automatic verification and augmentation follows these steps:

Preprocessing As most of the EGY entries in Tharwa are in the diacritized lemmatized form (97.7%), we first carry

[3] We also plan on extending this augmentation step to additional dialects of Arabic.

out a set of preprocessing procedures in order to clean, lemmatize and diacritize the Arabic side of both parallel data sets EGY-ENG and MSA-ENG to render the resources compatible. For the sake of consistency, the lemmatization step is replicated on the English data. The tool we use for processing Arabic is MADAMIRA $v1.0$ (Pasha et al., 2014), and for English we use the TreeTagger (Schmid, 1995). Table 4.4. illustrates the frequencies of types and tokens in each side of the lemmatized diacritized parallel data as well as the number of aligned parallel sentence pairs for each parallel corpus.

Parallel Data	Aligned Sentences	Arabic Words		English Words	
		Tokens	Types	Tokens	Types
MSA-ENG	2,820K	68,887K	180K	60,312K	252K
EGY-ENG	447K	3,682K	117K	3,746K	144K

Table 1: MSA-ENG and EGY-ENG parallel data statistics

Word Alignment The lemmatized-diacritized corpora with the corresponding ENG translations are word aligned using GIZA++ (Och and Ney, 2000) producing pairwise EGY-ENG and MSA-ENG word alignment files, respectively.

Pivoting on English gloss all triples in the form of EGY-ENG-MSA are extracted from both alignment files resulting from the previous step. We refer to this set of triples as *TransDict*. Table 2 shows total number of triples in TransDict along with number of MSA-ENG and EGY-ENG word alignments and the percentage of Tharwa matching on the ENG with each of these word pair alignments respectively. TransDict is extracted from intersection of both word pair alignments of these corpora. Similar triples EGY-MSA-ENG are extracted from the Tharwa entries and referred to as *TharwaDict*.

Parallel Corpus	Tuples	Tharwa Matched ENG %
MSA-ENG	64,885K	32.8%
EGY-ENG	4,173K	18.7%
Triples in TransDict		7,447K

Table 2: Total number of tuples extracted from the parallel corpora and percentage of MSA-ENG and EGY-ENG used to create TransDict along with total number of triples in TransDict

We compare TransDict and TharwaDict by pivoting on the EGY lemma entry as the primary key. Among the 7.4M entries in TransDict, 28K match fully (on EGY-MSA-ENG) with one of TharwaDict entries; 6.2M entries match on EGY and ENG but not MSA; 1M entries match on EGY and MSA but not ENG; and 181K entries match on EGY only. Additionally, approximately 75K TharwaDict entries have EGY lemmas that are not known in TharwaDict (regardless of whether the predicted ENG or MSA is in TharwaDict). The entries not fully matching TahrwaDict are good candidates for augmentation into Tharwa. We plan to use crowd sourcing to verify the quality of some of these entries and use the positive and negative examples to help us learn how to automate the process of filtering out noisy entries from the massive (7M) triples generated from the TransDict creation process in a manner similar to Saleh and Habash (2009).

4.5. Quality Control

A large portion of Tharwa is compiled and revised manually by professional linguists. However, it is necessary to make sure that errors are minimized and data backups are regularly maintained. Therefore, to guarantee the quality of Tharwa we employ two types of automatic quality control checks that help annotators minimize errors and data loss.

Version-Control Tharwa is version-controlled using SVN to backup new versions and to retrieve old versions where needed. We developed an interface between linguists/developers and the SVN tool for checking in updates to the SVN and checking out the latest version. The tool checks the new version before accepting. For example, if an annotator is assigned specific fields to revise such as POS and rationality, then they are only allowed to modify those specific attributes. If a violation occurs and the annotator modifies other attributes without checking it out officially, the SVN tool rejects the modification and produces a detailed report. This is particularly useful for preventing any unintended changes, and avoiding version conflicts.

Automatic Consistency Checks Regarding the EGY and MSA data content, we developed several automatic checks for detecting errors related to improbable spelling or diacritization. The following are some of the rules we used:

- All words must be fully diacritized
- Shadda (ـّ ~) cannot be followed by Sukuwn (ـْ *o*)
- Ta-Marbuta ة *p* must be preceded by ـَ *a*, ا *A*, or آ |
- ى *Y* must be preceded by ـَ *a*
- Tanween appears only in word-final position
- Vowel marks are not allowed in word-initial position

We also implement automatic checks on the ENG data ensuring that proper nouns[4] are capitalized and no spelling errors exist in the ENG data.

5. Tharwa Description and Statistics

All features provided are manually annotated and constantly and iteratively undergo quality checking procedures including guidelines, revision cycles and random sample testing to ensure robust quality with high inter-annotator agreement. The following are the linguistic features specified for lexical entries in Tharwa.

CODA This the diacritized conventional orthography lemma form of the EGY entries (Habash et al., 2012b). This field has a single entry in it. The number of entries in the Tharwa dictionary is 73,348. The following statistics show the level of overlap between the EGY entry and their MSA equivalent as defined in Section 2.3. These statistics are calculated on the lemma entries only amounting to ~51K entries. 33.5% of the entries are Identical (meaning and diacritized form) to MSA words, e.g. بخيل *baxiyl* 'miserly, cheap'; 14.4% are semantic cognates, modulo some regular homographic/homophonic variation with MSA, e.g., EGY اتكسّر *Aitokas~ar* and MSA تكسّر *takas~ar* 'become

[4] We assume that if an EGY is a proper noun, then its corresponding ENG is also a proper noun.

ID	EGY		POS	Root	Pattern	MSA		ENG
25	آدي	*|diy*	dem	-	-	هذا	*h‘*A*	this
3077	اتأجّل	*Aito>aj~il*	verb	*>jl*	`AitoC1aC2C2iC3`	تأخّر	*ta>ax~ar*	be postponed
10541	بايخ	*bAyix*	adj	*bwx*	`C1AC2iC3`	سخيف	*saxiyf*	silly
15539	ترباس	*tirobAs*	noun	*trbs*	`C1iC2oC3AC4`	مزلاج	*mizolAj*	latch
17578	جنينة	*jinaynap*	noun	*jnn*	`C1iC2ayC3ap`	حديقة	*Hadiyqap*	garden
19857	خلبصة	*xalobaSap*	vbn	*xlbS*	`C1aC2oC3aC4ap`	عربدة	*Earobadap*	raucous
20591	دكاكيني	*dakAkiyniy*	adv	*dkn*	`C1aC2AC2iyC3iy`	سرا	*sir~A*	secretly
21941	رقّاصة	*raq~ASap*	noun	*rqS*	`C1aC2C2AC3ap`	راقصة	*rAqiSap*	female dancer
23334	زرار	*zurAr*	noun	*zrr*	`C1uC2AC3`	زرّ	*zir~*	button
24754	سوّاق	*saw~Aq*	noun	*swq*	`C1aC2C2AC3`	سائق	*sA}iq*	driver
37891	مشغولات	*ma$oguwlAt*	noun	*$gl*	`maC1oC2uwC3At`	تحف	*tuHaf*	artifacts

Table 3: Example entries from Tharwa

broken'; 13.2% are homographs/homophones but with additional senses not in MSA, e.g., EGY حاجة *HAjap* and MSA شيء *$ay'* 'thing'; and, 38.9% are completely distinct EGY entries, e.g., EGY بس *bas* and MSA فقط *faqaT* 'only'.

EGY Variants This field lists alternative naturally occurring orthographic variants of the EGY CODA entries as obtained from their original sources BADAWI, and ECAL. This field can have multiple variants both diacritized and undiacritized, e.g., EGY entry كثير *kiviyr* 'many, a lot' (pronounced /kitiyr/) has the variant كتير *kitiyr*.

POS Tags We have two POS tag fields, one for EGY and one for MSA. The POS tags comprises 34 tags including verb, noun, adjective, adverb, particle, demonstrative, proper noun, and vbn (deverbal nouns).

EGY Word Type Each entry is marked as being a lemma or surface inflected form. Every surface form entry in Tharwa is linked to its lemma entry.

EGY Number, Gender and Rationality The semantic features number, gender and rationality. For more information, see (Habash, 2010; Alkuhlani and Habash, 2011).

EGY Root This is the root consonant radicals of the word before any derivation or inflection takes place. The root information is provided for both EGY and MSA entries, e.g. root: *ktb* 'writing-related', has 39 derived lemmas in Tharwa.

EGY Morphological Pattern This is the templatic structure of the word. We provide both the morphological (or deep) patterns and morpho-phonological (or surface) patterns for words, for both EGY lexical entries as well as the MSA equivalents, e.g. EGY: اكتتب *Aikotatab* 'subscribe' pattern: `AiC1otaC2aC3`.

MSA Equivalent This is the corresponding MSA word of the EGY entry. The MSA words are fully diacritized and are in the same morphological form of the EGY entry, e.g., EGY منكتب *minokitib* and MSA مكتوب *makotuwb* 'written'.

ENG Equivalent This is the equivalent translation into English.
Table 3 includes some example entries.

6. Uses of Tharwa in NLP Applications

Tharwa is used effectively within EGY processing tools. We list here two such tools.

Elissa is a DA-to-MSA machine translation system that exploits Tharwa's EGY-MSA correspondents together with other dialectal dictionaries to improve its translation process specifically by producing MSA paraphrases of the EGY words in the input Arabic sentence (Salloum and Habash, 2011; Salloum and Habash, 2013). This replacement process by Elissa as a preprocessing step in Arabic to English translation outperforms state-of-the-art performance from 37.2% BLEU (Papineni et al., 2002) points to 38.1% BLEU points on the dev set (a net 0.9% BLEU absolute improvement) and an improvement of 1.4% BLEU absolute on a blind test set.

AIDA is a tool for automatic identification of dialectal Arabic on both the token and sentence levels (Elfardy and Diab, 2012; Elfardy and Diab, 2013). For each word in a given sentence AIDA decides whether the word is EGY, MSA, Both (for both MSA and EGY), Named-Entity, or Unknown. It relies on MADAMIRA (Pasha et al., 2014) which relies on the CALIMA morphological analyzer which in turns relies on Tharwa dictionary and SAMA lexical databases coupled with language models to decide upon these classes. The AIDA system achieves a performance of 76.2% compared against a majority baseline for MSA of 51.4% and for EGY of 45.6%.

7. Conclusion & Future Plans

In this paper we present a three-way, large-scale lexicon Tharwa bridging Egyptian Arabic, Modern Standard Arabic and English. Tharwa provides rich and deep linguistic information for each entry and is the first comprehensive three way electronic resource for Dialectal Arabic, and hence can aid different NLP tasks. It was built and compiled in several steps and stages by combining the information from different resources including a paper dictionary and an electronic, corpus-based full form dictionary, in addition to the underlying lexica used in the morphological analyzers SAMA and CALIMA. We have implemented several measures for quality control through a controlled GUI with an underlying structured database, SVN technology, and automatic consistency check metrics. We also verify and augment tharwa by crowd sourcing techniques as well as leveraging triangulation through parallel corpora. This is in addition to multiple rounds of manual quality checks by native annotators.

For future work, we continue to work on the improvement of Tharwa quality and coverage, especially leveraging parallel resources and crowd sourcing. We plan to add corpus based example sentence usages for the entries to the lexical entries. We would like to expand Tharwa to account for several other dialects such as Levantine, Iraqi, Tunisian and Gulf Arabic.

Acknowledgments

We would like to thank Owen Rambow for helpful discussions and feedback. We would like to acknowledge the numerous annotators who helped building Tharwa. This work was supported by the Defense Advanced Research Projects Agency (DARPA) Contract No. HR0011-12-C-0014. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

8. References

Abo Bakr, H., Shaalan, K., and Ziedan, I. (2008). A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

Alkuhlani, S. and Habash, N. (2011). A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon, USA.

Badawi, E.-S. and Hinds, M. (1986). *A Dictionary of Egyptian Arabic*. Librairie du Liban.

Brustad, K. (2000). *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Elfardy, H. and Diab, M. (2012). Token level identification of linguistic code switching. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING), IIT Mumbai, India.*

Elfardy, H. and Diab, M. (2013). Sentence level dialect identification in Arabic. In *Proceedings of ACL, Sofia, Bulgaria.*

Ferguson, C. F. (1959). Diglossia. *Word*, 15(2):325–340.

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., Karins, K., Rowson, E., MacIntyre, R., Kingsbury, P., Graff, D., and McLemore, C. (1997). CALLHOME Egyptian Arabic Transcripts. In *Linguistic Data Consortium, Philadelphia.*

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic transliteration. In Soudi, A., Neumann, G., and van den Bosch, A., editors, *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, chapter 2, pages 15–22. Springer.

Habash, N., Eskander, R., and Hawwari, A. (2012a). A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

Habash, N., Diab, M., and Rabmow, O. (2012b). Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (2002). Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006). Developing and using a pilot dialectal Arabic treebank. In *LREC*, Genoa, Italy.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Morristown, NJ, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Saleh, I. and Habash, N. (2009). Automatic extraction of lemma-based bilingual dictionaries for morphologically rich languages. In *Third Workshop on Computational Approaches to Arabic Script-based Languages at the MT Summit XII, Ottawa, Canada.*

Salloum, W. and Habash, N. (2011). Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Salloum, W. and Habash, N. (2013). Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013).*

Schmid, H. (1995). Treetagger, a language independent part-of-speech tagger. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Spiro, S. (1895). *An Arabic-English Vocabulary of the Colloquial Arabic of Egypt*. Al-Mokattam printing office.

Spiro, S. (1987). *Arabic-English Dictionary of the Colloquial Arabic of Egypt*. Librairie Du Liban.