Developing a Tag-Set and Extracting the Morphological Lexicons to Build a Morphological Analyzer for Egyptian Arabic

Amany Fashwan a.fashwan@gmail.com
Sameh Alansary s.alansary@alexu.edu.egy
Linguistics and Phonetics Department, Faculty of Arts, Alexandria University, Egypt

Abstract

This paper sheds light on an in-progress work for building a morphological analyzer for Egyptian Arabic (EGY). To build such a tool, a tag-set schema is developed depending on a corpus of 527,000 EGY words covering different sources and genres. This tag-set schema is used in annotating about 318,940 words, morphologically, according to their contexts. Each annotated word is associated with its suitable prefix(s), original stem, tag, suffix(s), glossary, number, gender, definiteness, and conventional lemma and stem. These morphologically annotated words, in turns, are used in developing the proposed morphological analyzer where the morphological lexicons and the compatibility tables are extracted and tested. The system is compared with one of best EGY morphological analyzers; CALIMA.

1 Introduction

After the emergence of social media networks, and specially, after the Arab Spring revolutions, the data has become available everywhere. This led to have an increased attention in the field of Natural Language Processing (NLP) for Colloquial Arabic Dialects (CADs) where the adopted NLP tools for Modern Standard Arabic (MSA) are not suitable to process and understand them (Harrat, Meftouh, & Smaïli, 2017).

An important challenge for working on these dialects is to create morphological analyzers or tools that provide all possible analyses for a particular written word out of its context (Salloum & Habash, 2014) since it is an essential step in most NLP applications such as machine translation, information retrieval, text to speech, text categorization …etc. (Habash, Eskander, & Hawwari, 2012).

Morphological segmentation is the process of converting the surface form of a given word to its lexical form with additional grammatical information such as parts of speech, gender, and number (Joseph & Chang, 2012). In Morphological Analyzer (MA) tool, the morphemes along with their morphological information of a given word are provided for all its possible analyses out of its context.

This paper presents an in-progress work for building a morphological analyzer for Egyptian Arabic. To build such a tool, a Part-of-Speech (POS) tag-set schema is developed depending on different criteria to be used in annotating our corpus morphologically. The annotated data is used in detecting the different analysis solutions of each word, extracting the morphological lexicons and the compatibility tables to allow only valid morphological analysis solutions to be generated by the proposed morphological analyzer.

The rest of this paper is organized as follows. In Section 2, the related works are reviewed, then the used corpus and the process of developing the tag-set schema are discussed in Section 3. In Section 4, the proposed morphological analyzer and the prosses of the automatic extraction of the used morphological tables and the compatibility tables are discussed. The discussion of the system current status, coverage and evaluation are reviewed in Section 5. Finally, the discussion of conclusion and future work are listed in Section 6.

2 Related Works

Whereas there are many trials for defining tag-set schemas for MSA, for example, Khoja's Arabic Tag-set (Khoja, Garside, & Knowles, 2001; Khoja S. , 2003), ARBTAGS Tag-set (Alqrainy, 2008), and Penn Arabic Treebank (PATB) Part-of-Speech Tag-set (Maamouri & Bies, 2004), only few trials interested in EGY; (Maamouri, Krouna, Tabessi, Hamrouni, & Habash, 2012) who present a tag-set schema (ARZATB tag-set) that is based on the PATB guidelines (Maamouri M. , Bies, Krouna, Gaddeche, & Bouziri, 2009). They compare tags for Egyptian (ARZ) with those used in MSA. The

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 142 - 160

tags specify the forms of the morphemes used in constructing a word, but do not address discrepancies between morpheme form and functions. For example, the broken plural nouns in this tag set are treated the same as singular nouns: 'رجالة' /riggɑːl+æh/ 'men' is tagged as NOUN+NSUFF_FEM_SG. A new POS tag-set (CAMEL POS) is opted to be used in (Khalifa, et al., 2018). It is inspired by the ARZATB tag-set and guidelines. It is designed as single tag-set for both MSA and the dialects to facilitate research on adaptation between MSA and the dialects, support backward compatibility with previously annotated resources and enforce a functional morphology analysis that is deeper and more compatible with Arabic morphosyntactic rules than form-based analysis.

The lexicon and rules are the core knowledge base of any morphological analysis/generation system (Habash, 2010). The trials for modeling dialectal Arabic (DA) morphology have followed one of two directions. The first direction interested in extending MSA tools to cover dialectal phenomena. Some trials built their Egyptian colloquial lexicon for morphological analyzer on the top of Buckwalter Arabic Morphological Analyzer (BAMA) Version 2.0 (Buckwalter, 2004); (Shaalan, Bakr, & Ziedan, 2007), (Abo Bakr, Shaalan, & Ziedan, 2008), (Salloum & Habash, 2011), (Habash, Roth, Rambow, Eskander, & Tomeh, 2013), (Habash & Rambow, 2005), (Al-Sabbagh & Girju, 2010), (Diab, et al., 2014), (Maamouri M. , et al., 2006) and (Al Ameri & Shoufan, 2021). The second direction interested in modeling DA morphology directly; (Kilany, et al., 2002), (Habash & Rambow, 2006), (Habash, Eskander, & Hawwari, 2012), (Habash, Diab, & Rambow, 2012), (Mohamed, Mohit, & Oflazer, 2012), (Eskander, Habash, & Rambow, 2013), (Maamouri M. , et al., 2014), (Samih & Kallmeyer, 2017), (Zalmout, Erdmann, & Habash, 2018) and (Habash, Marzouk, Khairallah, & Khalifa, 2022).

Handling the problem of lacking standard orthography for colloquial Arabic dialects is very important for building the morphological analyzers. There are few works proposed the EGY to offer a set of orthographic rules, standards, and conventions for dialectal Arabic varieties; (Darwish, et al., 2018) is an attempt to conventionalize the orthography close to the dialectal pronunciation as much as possible regardless of the way a word is typically written. (Habash, Diab, & Rambow, 2012) provides detailed description of Conventional Orthography for Dialectal Arabic (CODA) as applied to EGY. A unified common set of guidelines and meta-guidelines that help in creating dialect specific conventions is presented in (Habash, et al., 2018) applied to 28 Arab city dialects including Cairo, Alexandria, and Aswan.

Lacking annotated resources considered as the bottleneck for processing and building robust tools and applications. However, low-resource languages still lack datasets, such as the Arabic language and its dialects. EGY has received a growing attention for building corpora that may be useful for many purposes such as dialect identification or sentiment analysis, for example, but only (Abo Bakr, Shaalan, & Ziedan, 2008), (Maamouri M. , et al., 2014), (Al-Sabbagh & Girju, 2012), (Bouamor, et al., 2018), and (Darwish, et al., 2018) are interested in building multi-dialect, multi-genre, morphologically annotated corpora that include EGY.

However, these annotated corpora have few shortcomings: none of them are freely available for use; they also do not represent enough variety of resource. Moreover, some of them normalize the orthography to MSA-like standards which fail to grasp the dialectal orthography differences, e.g., 'كتير' /kitiːr/ normalized as 'كثير' /kaθiːr/. Since MSA and the colloquials share a large proportion of their lexicon, the MSA tags are considered as much as possible as in (Maamouri, Krouna, Tabessi, Hamrouni, & Habash, 2012) and (Khalifa, et al., 2018). Nevertheless, we prefer to develop our own tag-set schema since we differ from (Maamouri, Krouna, Tabessi, Hamrouni, & Habash, 2012) in that we detect the tag according to its paradigmatic forms alongside its syntagmatic functions, as in (Khalifa, et al., 2018), as much as possible rather than depending on the morpheme form only. In addition, we opted to add more detailed tags in order to be more suitable to describe EGY, such as adverb of time, adverbs of place and adverbs of manner, and combine or split other tags that are described in the previously related-work tag-set schemas. Moreover, we feel the need to build a larger and more robust corpus, adding more various resources and genres. Consequently, this motivates us for building a new morphologically annotated resource for EGY to help in building the proposed morphological analyzer. It provides the conventional orthography

guidelines and develops a more suitable POS tag set for the EGY. For accessing our corpus, follow the link in[1].

3 The Corpus

The corpus used in developing the tag-set schema and the morphological analyzer consists of about 527,000 words representing about 82,700 tokens. The texts were selected from different sources such as social media, books and other web articles written in EGY (From Jan 2011- June 2019). In addition, these selected texts cover more than one genre. Lack of the standard orthography form in dialectal Arabic is handled by assigning for each word the conventional EGY Lemma and the conventional stem to be close to the EGY pronunciation as much as possible regardless the way a word is typically written. To improve the speed and accuracy of the manual morphological annotation, an interface is developed that allows the annotators to concentrate on the task of providing the best morphological analysis of each word according to its context. Six skilled linguistic annotators are trained to morphologically annotate the corpus. The conventional orthography guidelines, the annotation process and the inter-annotation agreement are reviewed in (Fashwan & Alansary, 2021).

3.1 The Morphological Features

The morphological annotation process includes adding features to a word in context, including its morphology, semantics, and other aspects. In the current used corpus, each document is saved in a database where there are several features that are added to each word. These features are: Raw Word, Edited Word, EGY Conventional Lemma, MSA Lemma, Person, Gender, Number, Definiteness, Gloss, POS tags and Conventional Stem.

The EGY Conventional Lemma is detected depending on the conventional orthography guidelines discussed in (Fashwan & Alansary, 2021). It is undiacritized, in this stage, due to the difference in the pronunciation among EGY sub-dialects, which is reflected in how a word may be diacritized, but, in the next stage, it is planned to be diacritized depending on one variety.

Not all EGY Lemmas have a corresponding MSA Lemma. For example, the origin of the word /mæʕæliʃ/ ‘معلش’ ‘sorry/excuse’ is /mɑ: ʕælæjhi ʃæjʔ/ ‘ما عليه شيء’. In this case, the MSA Lemma is assigned as combined ‘CMB’. It is worth mentioning that not all combined words are handled in the same manner; some words are split into more than one word assigned with their suitable POS tags according to the conventional orthography guidelines. Another case is the loanwords that are adopted in EGY and do not have MSA Lemma, for example, the word /niʃæjjær/ ‘نشير’ ‘we share’. In this case, the MSA Lemma is assigned as ‘LNW’. In addition, there are some words that are used in EGY, but its linguistic source is unknow. These words may have a counterpart meaning in MSA, consequently, the MSA Lemma is assigned. For example, the counterpart MSA lemma of the word /ʔiddæ:/ ‘إدى’ ‘give;provide’ is /ʔæʕṭæ:/ ‘أَعْطَى’.

The Gender takes two values: **1) “M”** for Masculine, or **2) “F”** for Feminine. The number takes one of four values: **1) “S”** for Singular, **2) “D”** for Dual, **3) “P”** for Plural, or **3) “B”** for Broken Plural /jæmʕ ɑt-tæksi:r/ ‘جمع التكسير’. The Definiteness takes one of three values: **1) “D”** for Definite, **2) “I”** for Indefinite, or **3) “E”** added through being the governor of an EDAFAH possessive construction /ʔiḍɑ:fæh/ ‘إضافة’.

The following sub-section defines the tag-set design schema used in assigning the suitable pos tag for all prefixes, suffixes, and stems in the compiled corpus.

3.2 The Tag-Set

The used POS tag-schema, in this work, specifies the suitable tags and sub-tags for prefixes, suffixes and stem. The current representation treats affixes and stems as separate tokens. It resembles the BAMA’s representation (Buckwalter, 2004; Habash, Eskander; Hawwari, 2012). Depending on the linguistic characteristics of EGY and general POS tag-set design criteria in (Atwell, 2008) such as mnemonic tag names, the underlying linguistic theory, classification by form or function, categorization problems, tokenization issues, …etc., there are several decisions are considered while defining the current POS tag-set design criteria:

- Since MSA and the Colloquials share a large proportion of their lexicon (Parkinson,

[1] https://forms.gle/3cpu1orvy4ohrosB9

1981), the MSA tags are considered as much as possible.

- The tag is intended to remain readable by linguists.
- The tag is detected according to its paradigmatic forms alongside its syntagmatic functions as much as possible.
- No 'combined tags' are used. Consequently, some words are needed to be split into their component morphemes where each morpheme is tagged separately.
- Since not all tags in MSA are suitable for the linguistic characteristics in EGY, more detailed compatible tags are needed.

In what follows, the POS of stem, prefixes and suffixes of the word are detailed in addition to its attributes:

1. **Stem:**

Nouns: The three main classified tags of MSA, namely: Noun, Verb, and Particle are applied in the current POS tag design schema. In (Al-Dahah, 1989), nouns are classified into 21 sub-classes, and other classifications overlap. In the current design schema, the noun is classified into 16 sub-classes. Appendix A provides a description of noun types as classified in the current proposed schema with their examples. In noun POS tags, the only tag that does not follow the Traditional Arabic Grammar is the adverb of degree.

Verbs: The verbs in Arabic are of two types: inflected and non-inflected. The inflected verb is classified, depending on its voice, into two types: active and passive. While active verb is classified, depending on the tense and the morphological forms, into three groups: Perfect Verb **(PV)**, Imperfect Verb **(IV)** and Imperative Verb (RV), the passive verb is classified into two groups only: Perfect Verbs **(PV)** and Imperfect Verbs **(IV)**. The non-inflected verbs, also known as non-conjugated verbs, appear in perfect, imperfect, or imperative form. In the current design schema, the verbs are classified into four sub-classes as Appendix B shows. Three types are defined depending on the classical Arabic classification and only the Pseudo Verb tag is defined depending on the linguistic nature of EGY texts.

Particles: They are words that do not belong to nouns or verbs, but they add specific meaning to them in a sentence or connect two or more sentences. In traditional Arabic, the particles may also be classified into two groups according to their effect on nouns or verbs. The governing particles /ɑl-ħuru:f ɑl-ʕɑ:milæh/ 'الحروف العاملة' that affect the form of the following noun or verb; and the non-governing particles /ɑl-ħuru:f ɣæjr ɑl-ʕɑ:milæh/ 'الحروف غير العاملة' which do not affect the form of the following noun or verb (Al-Dahah, 1989). Appendix C indicates how particles are defined and classified in EGY.

Others (Residual): Others (residuals) include foreign words, non-Arabic words, punctuation marks, Emojis, abbreviations, numbers, in addition to words that express the speaker's reaction to a particular suggestion or sentence. E.g., /hhhh/ 'هههههه', /ti:t/ 'تيت' and /jöh/ 'يوه' as Appendix D shows.

2. **Prefixes:**

In the current design schema, the prefixes are defined depending on the previously described stems particles in addition to newly defined tags, as Appendix E indicates. As concerning to imperfect and imperative particles, information about verb person, gender, and number **(PGN)** of the verb subject are added since these particles are represented in prefixes for imperfect and imperative verbs only.

3. **Suffixes:**

Two types of suffixes tags are defined depending on the previously described tags of stem. In addition, the noun's suffix inflections are defined where the nouns may be inflected for suffixes of person, gender, definiteness, number such as 'ين' /i:n/, 'ات' /ɑ:t/, 'ة' /t/, etc. They are given the tag 'NSUF' alongside their gender, number, and definiteness **(GND)**. It is worth mentioning that the same suffix may be attached with different gender, number, or definiteness since we detect the tag according to its functions rather than its form. For example, the 'ة' /t/ 'taa marbouta' may be assigned 'NSUF_FS' as in 'مدرسة' /mædræsæ/ 'school', 'NSUF_MS' as in 'أسامة' /ʔusɑ:mæ/ 'Osama', 'NSUF_MB' 'رجالة' /riggɑ:læ/ 'men;people', etc. In case the noun is not inflected for suffix as in 'ولد' /wælæd/, a word is given 'null/NSUF' in POS annotation alongside its stem's **(GND)**.

The verb inflections are represented in suffixes for all verb tenses and information about verb person, gender, and number **(PGN)** of the verb subject are added.

Since the case endings are dropped out in EGY writing except the case morpheme 'ا' /ʔælif ɑt-

tænwi:n/ ‘ألف التنوين’ ‘Alif for nunnation’ that may be written in some words, for example, /ʃukræn/ ‘شكرا’ ‘thanks’, /giddæn/ ‘جدا’ ‘very much’, and /mæsælæn/ ‘مثلا’ ‘for example’, there is a need to add a tag that represents this information, although it is a syntactic rather than morphological. Consequently, the tag ‘CASE’ is added to the previous enclitic tags. For more details about used suffixes in the current POS tag-set schema, check as Appendix F.

3.3 Corpus Annotation Current State

As a first step, about 318,940 words are annotated morphologically. These annotated words are the milestone for the automatic extraction of the morphological lexicons and the compatibility tables used for developing the proposed morphological analyzer. They are also planned to be used to extend the annotation to the remaining words of the EGY corpus, automatically. Table 1 shows the frequencies of POS tags in the currently annotated corpus. After the residuals that are annotated in the whole corpus data, the most frequent tags in the corpus are the nominals (NOU, NOU_NUM, NOU_SUP, and NOU_PRP).

Tag	Frequency
Others (Residuals)	72,330
Nouns (NOU, NOU_NUM, NOU_SUP and NOU_PRP)	81,981
Prepositions (PRP)	33,345
Pronouns (PRN, PRN_DEM and PRN_REL)	29,880
Verbs (VER_ACT, VER_PSV, VER_DFC and VER_SUD)	24,658
Other Particles (PRT_NEG, PRT_FUT, PRT_VER, PRT_INT, PRT_VOC, PRT_AUG, PRT_EXC and PRT_EMP)	18,629
Adverbs (ADV_PLC, ADV_TIM, ADV_TPL and ADV_DGR)	15,453
Adjectives (ADJ, ADJ_SUP and ADJ_NOM)	13,790
Conjunctions (CNJ and CNJ_SUB)	9,659
Interrogative Pronouns (PRN_INT)	4,847

Table 1: POS Tag Frequencies.

The annotated data contains about 12,100 unique conventional EGY lemmas representing about 18,400 MSA lemmas. Each EGY lemma is associated with different stems and each stem is associated with their different tags and conventional stems according to their contexts.

4 The Morphological Analyzer

EGY Arabic words are rarely written with diacritic marks; consequently, they may have many morphological analyses, and the number of these analyses differs from one word to another. Since the morphological analyzer deals with words out of their contexts, it should be able to produce all possible analyses of each form, identify the part-of-speech of each analysis solution of the word (i.e., noun, verb, and particle) and identify the morphological features (i.e., gender, number, time, and person). It is not an easy task to capture all analysis solutions of each word, but the annotated corpora one of the most important resources that can be helpful in detecting these solutions depending on the different contexts of the same word.

We follow a concatenative lexicon-driven approach for the annotation of our morphological corpus. The concatenation can be defined as a sequence of prefix(es), stem and suffix(es) or as a sequence of proclitic(s), word form and enclitic(s), where the morphological segments are recognized and processed as part of the annotation process. We adopt the former scheme, where the plan is to allow for the conversion between the two in our morphological analyzer.

The focus in this paper is on the prefix(es), stem and suffix(es) representation. Our approach resembles the adopted one in Buckwalter Version 2.0 (Buckwalter, 2004) who uses a simple prefix-stem-suffix representation where the stem is used as the base form and morphotactics and orthographic rules are built directly into the lexicon itself instead of being specified in terms of general rules that interact to realize the output. It has three components: the lexicon, the compatibility tables, and the analysis engine.

4.1 Extracting the Morphological Lexicons and Compatibility Tables

These lexicons need to meet certain specifications such as high coverage, high level of quality, directly reusable in NLP tools, and freely available to potential users (Sawalha, 2011). The morphological lexicons are essential for generating all possible combinations of morphemes. The wrong combinations of morphemes of lexicons are

the major problem of generation. Consequently, the compatibility tables are needed for filtering out these wrong combinations.

The unique solutions of the morphologically annotated words in our corpus are used to automatically generate the morphological lexicons: the prefixes lexicon (dictPrefixes), the stems lexicon (dictStems), the suffixes lexicon (dictSuffixes) and the out of vocabulary (OOV) lexicon (dictOOV). In addition, the compatibility tables combAC, combAC and combBC are extracted to help in obtaining the valid concatenations among the different morphological categories of Prefixes, Stem and Suffixes lexicons.

For extracting these lexicons and the compatibility tables, we start with the unique annotated solutions in our corpus as a combination of (EGY lemma, MSA Lemma, conventional stem, [prefix+stem+suffix] and features). Figure 1 shows the process for extracting the features needed for building the lexicons from these solutions.

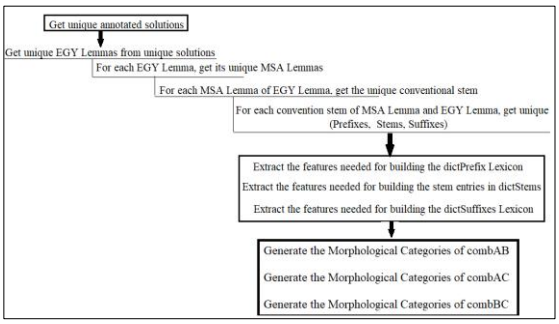

Figure 1: Lexicons and Compatibility Tables Extraction Process.

The extracted information in each Lexicon are as follows:

1. **Stems Lexicon (dictStem)**

In this lexicon, one of three keys appears at the beginning of each line to represent a specific morphological feature while parsing the stems lexicon. These keys are as follows:

- **';; '**: what follows this key represents the conventional EGY lemma for the subsequent lines till the next ';; ' key.

- **';;- '**: what follows this key represents the MSA diacritized lemma for the subsequent lines till the next ';;- ' key. The same EGY lemma may have a different MSA lemma due to the different diacritization of the MSA lemma.

- **';;-- '**: it is the default key for representing the stem entry and its conventional orthography. If it is the only written key, then the stem entry, the output stem, and the conventional stem are the same. If there are other keys found within the line after it, this means that there are more details while handling the stem and its conventional orthography. This helps in handling many processes such as transformation, omission, and assimilation that occur for the analyzed words. For Example, the '^' key may appear within the line after ';;-- ' key, then the word before it may represent different morphological information. For example, it could represent a stem's morpho-phonological changes due to the assimilation when the word is attached to an enclitic: (;;-على ع^على /ʕæl^ʕælæ:/ 'on;above') as in the word 'علي' /ʕælæj+jæ/ 'on me' where the stem ends with /æ:/ 'ى' and the enclitic begins with /j/ 'ي', which leads to an assimilation process where the two conceding /æ:j/ 'يي' are transformed to /jj/ 'يّ'.

- When none of the previous keys appear at the beginning of the input line, a line is parsed as it consists of three tab-delimited fields: **1)** the morphological category that controls the compatibility of prefixes-stem-suffixes, **2)** the English gloss(es) of stem in addition to information about the number, gender, and definiteness (in case the stem is a noun, adjective or adverb), or the person of the stem (in case of verbs only and pronouns) and **3)** the selective POS tags that appear in the analysis output. The morphological category of each stem is extracted automatically depending on the suffixes that are attached to each solution. For example, the "N-ap-I" category refers to the indefinite nouns that are attached to "ة/NSUF_(GN)I" as in "مدرسة" /mædræsæ/ 'school' and the "IV-y-0" category refers to the "ي/IVSUF_2F" suffix that is not attached to another suffix as in "تبتسمي" /ti-btisim+i:/ 'you + smile'.

2. **Prefixes and Suffixes Lexicons (dictPrefixes) and (dictSuffixes)**

In these lexicons, all used prefixes and suffixes of the annotated words are listed. They consist of four tab-delimited fields: **1)** the prefix/suffix entry

in Arabic orthography without any diacritics, **2)** the morphological category that controls the compatibility of prefixes-stems-suffixes, **3)** the English gloss(es) of each prefix/suffix part in the prefix/suffix entry, and **4)** the selective POS tags that appear in the analysis output. The morphological category of each prefix in the corpus is detected automatically depending on the prefixes' parts in addition to the tag of the stem that is attached to them. For example, the 'IVPrf-wa-bi-n' category represents the 'و/CNJ' prefix in addition to progressive particle 'ب/PRT_PRG' and 1st person plural prefix that are attached to Imperfect Verbs 'ن/IVPRF_1P' as in 'وبنرسم' /wibni-rsim/ 'and + we + draw;trace;sketch'. The morphological category of each suffix in the corpus is detected automatically depending on the suffixes' parts in addition to the tag of the stem that is attached to them. For example, the 'ADSuf-nl-h' category represents the 3rd person pronouns that may be attached to the adverbs as in 'بينه' /beɪn+uh/ 'between;among + him'.

3. Out of Vocabulary Lexicon (dictOOV):

This lexicon is created to be used in predicting the OOV words. It consists of three tab-delimited fields: **1)** the unique stem patterns, **2)** the morphological category that controls the compatibility of prefixes-stems-suffixes and **3)** the selective POS tags that appear in the analysis output. For detecting the stem patten of each stem, the consonants are represented by the placeholder "-", while weak letters 'حروف العلة' /ħuruːf ɑl-ʕillæh/ and hamazat ('أ', 'إ', 'ؤ', 'ئ', 'ء') are kept as they are. For example, the stem pattern of 'اعمل' /iʕmil/ 'do;act;make', 'اهرب' /ihrab/ 'run away' and 'اكتب' /iktib/ '' is '---ا'.

4. The Compatibility Tables

The compatibility table (combAB) lists the two compatible morphological categories of Prefixes and Stems. It consists of two tab-delimited fields: **1)** Prefix Morphological Category and **2)** Stem Morphological Category that appear together in the annotated data. The compatibility table (combAC) lists the two compatible morphological categories of Prefixes and Suffixes. It consists of two tab-delimited fields: **1)** Prefix Morphological Category and **2)** Suffix Morphological Category that appear together in the annotated data. The compatibility table (combBC) lists the two compatible morphological categories of Stems and Suffixes. It consists of two tab-delimited fields: **1)** Stem Morphological Category and **2)** Suffix Morphological Category that appear together in the annotated data. The morphological categories that are not listed in the compatibility tables are simply incompatible.

4.2 The Analyzer

The current morphological analyzer goes through four main steps to get all possible morphological analyses of the input words:

1) Text Preprocessing and Lexicons Parsing: in this step, it is important to detect the word boundaries of the input text since it is essential step for the word segmentation process. In addition, the 'dictPrefixes' and 'dictSuffixes' lexicons are parsed to get the four tab-delimited fields in dictionaries where the prefix/suffix entry is the default key for these dictionaries. Each line in 'dictStems' lexicon is parsed in different manner depending on the key used at the beginning of each line as mentioned above (section 4.1). The conventional stem in this lexicon is handled to get all possible stem variations of the input word. For example, the stem variations ('إلي', 'إلي', 'إلى', 'ألى', 'ألي', 'ألى', …etc.) are generated automatically from the conventional stem 'إلى' /ʔilæː/ 'to;towards' to avoid writing all these expected stem variations in the lexicon. The stem variations that cannot be predicted automatically are added to the 'dictStems' lexicon with their suitable morphological category.

2) Word Segmentations and Compatibility Check: For suggesting different segmentations of the same word, the dictionaries of the parsed lexicons are used. The three morphological categories of the three components are checked in the compatibly tables as figure 2 shows. If they are found together, then they are compatible, and this is a valid solution. Else, they are incompatible, and this is not valid solution.

3) Dealing with OOV Words: For handling the OOV words, the analyzer tries, first, to split the input word depending on its beginning and end. For Example, it splits OOV word that begin with /jɑː/ 'يا' since attaching it to another word is a common spelling mistake in EGY writings as in 'يارب' /jɑː ræbb/ 'Oh, Lord' and 'ياسلام' /jɑː sælɑːm/ 'really'. To keep the original word and the split words in output analysis, another feature is added; normalized word 'norm_word'. All possible solutions for each part are detected regardless of the solutions of the two parts are compatible according to their context or not. In case there is no

rule for splitting the OOV word or it is split but only one of its parts has analysis solution, the analyzer tries to detect the prefix and the suffix of the input word. If they are predicted, the stem is converted to its corresponding pattern as mentioned above. If the stem pattern is found in the 'dictOOV', the morphological categories of prefix, suffix, and the suggested stem pattern are checked in the compatibly tables. If they are found together, then they are compatible, and this is a valid solution, but no lemma or gloss are detected.

4) Output Solutions: After getting all possible solutions of the input text for all words and handling the OOV words, the output valid solutions are saved in XML format.

5 The Current Status

The extracted morphological stem lexicon contains 39K stems corresponding to about 12,100 EGY Lemmas and about 18,400 MSA lemmas. The extracted prefixes and suffixes lexicons contain

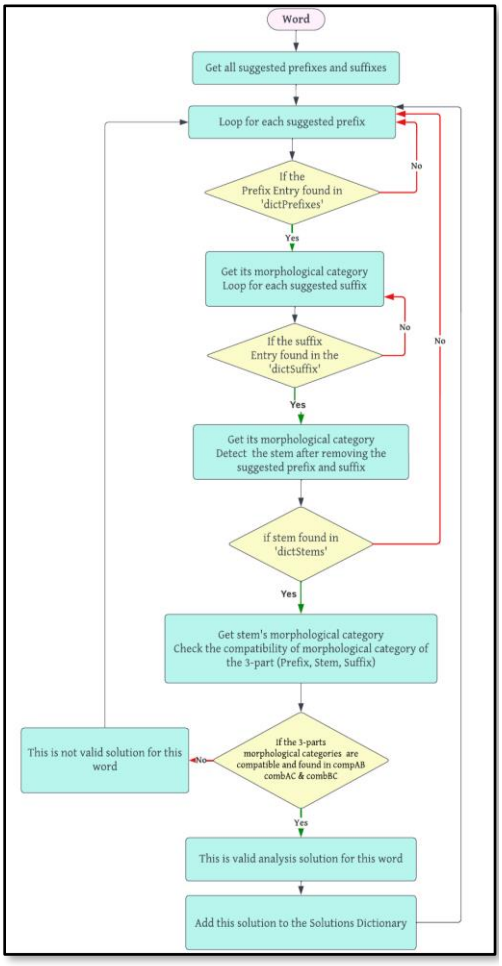

Figure 2: Workflow for Suggesting Words' Segmentations and Get Valid Solutions.

324 complex prefixes and 661 complex suffixes (unique undiacritized form and POS tag combinations). Since the annotation process of our corpus is still in progress, the covered stems, prefixes, suffixes and lemmas are still limited compared to CALIMA analyzer (Habash, Eskander, & Hawwari, 2012) that has 100K stems corresponding to 36K lemmas in addition to 2,421 complex prefixes and 1,179 complex suffixes (unique diacritized form and POS tag combinations).

5.1 Coverage Evaluation

We tested our analyzer against a sample of our manually annotated EGY corpus of 5,000 words which was not used as part of its development, i.e., a completely blind test. This evaluation is a POS recall evaluation. It is not about selecting the correct POS answer in context. We do not consider whether the EGY lemma or the MSA Lemma choice are correct or not. We compare our system results with CALIMA coverage. The results are reported in Table 2. The 'Correct Answer' column indicates the percentage of the test words whose correct analysis in context appears among the analyses returned by the analyzer. The 'No Correct Answer' column presents the percentage of time one or more analyses are returned, but none matching the correct answer. The 'No Analysis' column indicates the percentage of words returning no analyses.

	Correct Answer	**No Correct Answer**	**OOV**
Our System	66.9%	10.3%	22.8%
CLIMA	82.1%	9.6%	8.3%

Table 2: Comparing Results with CALIMA.

6 Conclusion and Future Work

The POS tag-set schema is developed and about 318,940 words are morphologically annotated, and the morphological lexicons and the compatibility tables are automatically extracted. The analyzer output is compared to CALIMA output. We plan to make this tool public so it can be used by other people working on EGY NLP tasks, from annotating corpora to building morphological disambiguation tools. To enhance our results, we plan to continue improving the coverage of our

analyzer using a variety of methods. First, we are investigating techniques to automatically fill in the tag categories gaps using information from multiple entries in our annotated corpus belonging to different lemmas that share similar characteristics, e.g., hollow verbs. Another direction is to increase the stems entries by checking stems, in BAMA's stems lexicon, for those words that are common between EGY and MSA and adapting their morphological category to be more suitable for EGY. Furthermore, we plan to add additional features such as the diacritized EGY lemmas and the diacritized stems.

References

Abo Bakr, H. A., Shaalan, K., & 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. Retrieved 7 27, 2018, from https://www.researchgate.net/profile/Khaled_Shaalan/publication/206006074_A_Hybrid_Approach_for_Converting_Written_Egyptian_Colloquial_Dialect_into_Diacritized_Arabic/links/0912f505a298ee3308000000.pdf

Al Ameri, S. S., & Shoufan, A. (2021). Building Lexical Resources for Dialectical Arabic. In *Natural Language Processing for Global and Local Business* (pp. 332-364). IGI Global.

Al-Dahah, A. (1989). *A Dictionary of Arabic Grammer in Charts and Tables " معجم قواعد اللغة العربية في جداول ولوحات".* Beirut, Lebanon: Librairie du Liban publisher.

Alqrainy, S. (2008). A morphological-syntactical analysis approach for Arabic textual tagging. Leicester, UK: De Montfort University. Retrieved 11 8, 2021

Al-Sabbagh, R., & Girju, R. (2010). Mining the Web for the Induction of a Dialectical Arabic Lexicon. *In LREC.* Retrieved 6 23, 2019, from https://www.researchgate.net/profile/Rania_Al-Sabbagh/publication/220746429_Mining_the_Web_for_the_Induction_of_a_Dialectical_Arabic_Lexicon/links/02e7e51597e45d9c5d000000.pdf

Al-Sabbagh, R., & Girju, R. (2012). YADAC: Yet another Dialectal Arabic Corpus. *In LREC*, (pp. 2882-2889). Retrieved 6 23, 2019, from https://pdfs.semanticscholar.org/67ad/6967eb602c7416e8aaa138bb4c45a23b4e07.pdf

Atwell, E. (2008). Development of tag sets part-of-speech tagging. In A. Ludeling, & M. Kyto (Eds.), *Handbook, Corpus Linguistics: An International* (Vol. 1, pp. 501 - 526). Walter de Gruyter. Retrieved 7 5, 2019, from https://eprints.whiterose.ac.uk/81781/1/DevelopmentTagSetPOSTagging.pdf

Bouamor, H, Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. *In Proceedings of the 11th International Conference on Language Resources and Evaluation.* Retrieved 8 7, 2018, from http://www.lrec-conf.org/proceedings/lrec2018/pdf/351.pdf

Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0.* Linguistic Data Consortium, University of Pennsylvania, 2004. LDC Catalog No.: LDC2004L02. Retrieved 11 20, 2019, from https://catalog.ldc.upenn.edu/LDC2004L02

Darwish, K., Mubarak, H., Abdelali, A., Eldesouki, M., Samih, Y., Alharbi, R., Attia, M., Magdy, W., & Kallmeyer, L. (2018). Multi-Dialect Arabic POS Tagging: A CRF Approach. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC.* Retrieved 6 19, 2019, from https://www.aclweb.org/anthology/L18-1015

Diab, M., Al-Badrashiny, M., Aminian, M., Attia, M., Dasigi, P., Elfardy, H., Eskander, E., Nizar, H., Hawwari, A., & Salloum, W. (2014). Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. *In LREC*, (pp. 3782-3789). Retrieved 7 28, 2018, from https://www.researchgate.net/profile/Mohammed_Attia2/publication/305489865_Tharwa_A_Large_Scale_Dialectal_Arabic-Standard_Arabic-English_Lexicon/links/582f14bb08ae138f1c034db8.pdf

Eldesouki, M., Samih, Y., Abdelali, A., Attia, M., Mubarak, H., Darwish, K., & Laura, K. (2017). Arabic multi-dialect segmentation: bi-LSTM-CRF vs. SVM. Retrieved 9 1, 2022, from https://arxiv.org/pdf/1708.05891.pdf

Eskander, R., Habash, N., & Rambow, O. (2013). Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora. *In Proceedings of the 2013 conference on empirical methods in natural language processing*, (pp. 1032-

1043). Retrieved 7 4, 2019, from https://www.aclweb.org/anthology/D13-1105

Fashwan, A., & Alansary, S. (2021). A Morphologically Annotated Corpus and a Morphological Analyzer for Egyptian Arabic. Procedia Computer Science. *189*, pp. 203-210.

Habash, N. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies Journal, Vol. 3*, pp. 1-187.

Habash, N., & Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. *In Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, (pp. 573-580). Retrieved 7 4, 2019, from https://www.aclweb.org/anthology/P05-1071

Habash, N., & Rambow, O. (2006). MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, (pp. 681-688). Retrieved 6 24, 2019, from https://www.aclweb.org/anthology/P06-1086

Habash, N., Diab, M., & Rambow, O. (2012). Conventional Orthography for Dialectal Arabic. *In LREC*, (pp. 711-718). Retrieved 6 19, 2019, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/579_Paper.pdf

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., . . . Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography., (pp. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)). Retrieved 6 19, 2019, from https://www.aclweb.org/anthology/L18-1574

Habash, N., Eskander, R., & Hawwari, A. (2012). A morphological Analyzer for Egyptian Arabic. *In Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology* (pp. 1-9). Association for Computational Linguistics. Retrieved 7 29, 2018, from https://aclanthology.org/W12-2301.pdf

Habash, N., Marzouk, R., Khairallah, C., & Khalifa, S. (2022). Morphotactic Modeling in an Open-source Multi-dialectal Arabic Morphological Analyzer and Generator. *In Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, (pp. 92-102). Retrieved 7 1, 2022, from https://aclanthology.org/2022.sigmorphon-1.10.pdf

Habash, N., Roth, R., Rambow, O., Eskander, R., & Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal Arabic. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 426-432). Retrieved 7 3, 2019, from https://www.aclweb.org/anthology/N13-1044

Harrat, S., Meftouh, K., & Smaïli, K. (2017). Machine translation for Arabic dialects (survey). *Information Processing and Management*. Retrieved 7 28, 2018, from https://www.researchgate.net/profile/Kamel_Smaili/publication/319423437_Machine_translation_for_Arabic_dialects_survey/links/5a0a32f0a6fdcc2736dea63b/Machine-translation-for-Arabic-dialects-survey.pdf

Chang, J. Z., & Chang, J. S. (2012). Word root finder: a morphological segmentor based on CRF. *Proceedings of COLING 2012: Demonstration Papers.* Retrieved from 2012: https://aclanthology.org/C12-3007.pdf

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., & Al Kaabi, M. (2018). A morphologically annotated corpus of Emirati Arabic. *In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Retrieved 5 29, 2021, from https://aclanthology.org/L18-1607.pdf

Khoja, S. (2003). APT: An automatic arabic part-of-speech tagger. *Doctoral dissertation*. Lancaster University. Retrieved 11 8, 2021

Khoja, S., Garside, R., & Knowles, G. (2001). An Arabic tagset for the morphosyntactic tagging of Arabic. *Doctoral dissertation, 13*, . Lancaster University (UK). Retrieved 8 11, 2021

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., & McLemore, C. (2002). *Egyptian Colloquial Arabic Lexicon.* LDC catalog number LDC99L22. Retrieved 8 5, 2019, from https://catalog.ldc.upenn.edu/LDC99L22

Maamouri, M., & Bies, A. (2004). Developing an Arabic treebank: Methods, guidelines, procedures, and tools. *In Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages*, (pp. 2-9). Retrieved 11 8, 2021, from https://aclanthology.org/W04-1602.pdf

Maamouri, M., Bies, A., Buckwalter, T., Diab, M. T., Habash, N., Rambow, O., & Tabessi, D. (2006). Developing and Using a Pilot Dialectal Arabic Treebank. *In LREC*, (pp. 443-448). Retrieved 8 4, 2018, from https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2006-developing-using-dialectal-arabic-treebank.pdf

Maamouri, M., Bies, A., Krouna, S., Gaddeche, F., & Bouziri, B. . (2009). *Penn Arabic treebank guidelines.* Linguistic Data Consortium.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., & Eskander, R. (2014). Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. *LREC*, (pp. 2348-2354). Retrieved 6 1, 2021, from http://www.lrec-conf.org/proceedings/lrec2014/pdf/1145_Paper.pdf

Maamouri, M., Krouna, S., Tabessi, D., Hamrouni, N., & Habash, N. (2012). *Arabic Treebanking Egyptian Arabic (ARZ) Morphological (ARZGM) Version 1.21 (With Permession).* Retrieved 08 21, 2022, from https://www.researchgate.net/publication/331315455_Egyptian_Arabic_Morphological_Annotation_Guidelines/stats

Mohamed, E., Mohit, B., & Oflazer, K. (2012). Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic. *In LREC*, (pp. 873-877). Retrieved 7 4, 2019, from http://nlp.qatar.cmu.edu/papers/465_Paper.pdf

Parkinson, D. B. (1981). VSO to SVO in Modern Standard Arabic: A study in diglossia syntax. In *al-'Arabiyya* (Vol. 14, pp. 24-37). Georgetown University Press. Retrieved 5 15, 2021

Salloum, W., & 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* (pp. 10-21). Association for Computational Linguistics. Retrieved 7 26, 2018, from http://delivery.acm.org/10.1145/2150000/2140535/p10-salloum.pdf?ip=196.204.161.40&id=2140535&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1560078805_b75acb510a9652c29cada891c1ec3bd2

Salloum, W., & Habash, N. (2014). ADAM: Analyzer for Dialectal Arabic Morphology. *Journal of King Saud University-Computer and Information Sciences, 26(4)*, 372-378. Retrieved 7 30, 2018, from https://www.sciencedirect.com/science/article/pii/S1319157814000342

Samih, Y., & Kallmeyer, L. (2017). Dialectal Arabic Processing Using Deep Learning. *Doctoral dissertation, Ph. D. thesis, Heinrich-Heine-Universität Düsseldorf, Düsseldorf.* Germany. Retrieved 8 5, 2018, from https://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-47492/Dissertation_younes_samih.pdf

Sawalha, M. S. (2011). *Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora.* University of Leeds.

Shaalan, K., Bakr, H., & Ziedan, I. (2007). Transferring egyptian colloquial dialect into modern standard arabic. *In International Conference on Recent Advances in Natural Language Processing (RANLP–2007)*, (pp. 525-529). Borovets, Bulgaria. Retrieved 6 10, 2019, from http://www.linguisticsnetwork.com/wp-content/uploads/Transferring-Egyptian-Colloquial-Dialect-into-Modern-Standard-Arabic-2.compressed.pdf

Zalmout, N., Erdmann, A., & Habash, N. (2018). Noise-Robust Morphological Disambiguation for Dialectal Arabic. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers), Volume 1* , pp. 953-964. Retrieved 7 4, 2019, from https://www.aclweb.org/anthology/N18-1087

A Nouns

Noun Sub Classes	Description and Example
Noun 'الاسم' /al-ism/ **(NOU)**	It is the common noun that refers to entities and concepts that have a more general reference than sub-tags. والولد راح مدرسته مع مامته في أول يوم وكل المدرسين رحبوا بيه. wil-wælæd ra:ħ mædræstu mæʕæ ma:mtu fi: ʔæwwil jöm wi-kull al-mudarrisi:n ræħæbu: bi:h
Proper Noun 'اسم العلم' /ism al-ʕælæm/ **(NOU_PRP)**	It is a noun that has a unique referential meaning in a context that is mutually exclusive with other entities. It refers to names of people, geographical entities, months, and acronyms. محمد هو اللي قالي الكلام ده أول شهر مارس كدة وكنا وقتها في إسكندرية. mæħæmmæd huwwæ illi: ʔa:lli: al-kæla:m dæh fi: ʔæwwil ʃæhr ma:ris kidæ wi-kunna: wæʔtæhæ fi: ʔiskindirijjæ
Numeral Noun 'اسم العدد' /ism al-ʕædæd/ **(NOU_NUM)**	It is a noun that indicates the quantity and order of countable nouns by transferring the numbers into the correct form of Arabic words. الأهلي غلب الزمالك واحد/صفر. al-ʔæhli: ɣælæb az-zæma:lik wa:ħid/ṣifr
Adjective 'الصفة' /aṣ-ṣifæ/ **(ADJ)**	It is a noun that describes or clarifies the meaning of the immediately preceding noun. البنت الشاطرة تسمع كلام مامتها. al-bint aʃ-ʃa:træ tismæʕ kæla:m ma:mitha:
Numeral Adjective 'الصفة العدد' /aṣ-ṣifæ al-ʕædæd/ **(ADJ_NUM)**	It is an adjective that indicates the quantity and order of countable nouns by transferring the numbers into the correct form of Arabic words. الحكاية الحادية عشر. al-ħika:jæh al-ħa:dijætæ ʕæʃær
Nominal Adjective 'الصفة الاسمية' /aṣ-ṣifæ al-ismijjæ/ **(ADJ_NOM)**	It is a noun that describes or clarifies the meaning of a noun, but it appears as the main predicate of a nominal phrase in the sentence. كانت حقيقي جميلة وعمري ما شفت بنت في أخلاقها أبدا. ka:nit ħæʔi:ʔi: gæmi:læ wi-ʕumri: ma: ʃuft bint fi: ʔæɣla:ʔha:
Superlative Adjective 'صفة تفضيل' /ṣifæt tæfḍi:l/ **(ADJ_SUP)**	It is a noun that is used for the comparative and superlative when comparing persons or things. It describes the immediately preceding noun. الحاجة الأحلى اللي تعملها إنك تسمع الكلام وإنت ساكت. al-ħa:gæ al-ʔæħlæ: illi: tiʕmilha: ʔinnak tismæʕ al-kæla:m wi-ʔintæ sa:kit
Superlative Noun 'اسم تفضيل' /ism tæfḍi:l/ **(NOU_SUP)**	It is a noun that is used for the comparative and superlative when comparing persons or things, but it appears as the main predicate of a nominal phrase in the sentence. والله دي كانت أجمل حاجة حصلتلي في حياتي. wal-læhi: di: ka:nt ʔægmæl ħa:gæ ħæṣælitli: fi: ħæja:ti:
Adverb of Place 'اسم المكان' /ism al-mæka:n/ **(ADV_PLC)**	It is a noun that indicates where the action of a verb is or was carried out. قعد جوة البيت طول اليوم. ʔæʕæd juwwæ al-beɪt ṭu:l al-jöm

Noun Sub Classes	Description and Example
Adverb of Time ‘اسم الزمان’ /ism ɑz-zæmɑ:n/ **(ADV_TIM)**	It is a noun that indicates when the action of a verb happened. It expresses a point in time, and it can also indicates how long something lasted or lasts. والله لسه شايفه إمبارح ومش عارف حقدر أشوفه تاني بكرة ولا لأ. wɑl-læhi: lissæ ʃɑ:jfuh ʔimbɑ:riħ wi-miʃ ʕɑ:rif ħæ-ʔdær ʔæʃu:fuh tɑ:ni: bukræ wællæ læʔ
Adverb of both Time and Place ‘اسم زمان ومكان’ /ism mækɑ:n wɑ-zæmɑ:n/ **(ADV_TPL)**	It is a noun that could be used as an adverb of time or place according to its context. كان رايح عند السوبر ماركت يجيب حاجات للبيت. kɑ:n rɑ:jiħ ʕænd ɑs-su:bær mɑ:rkit jigi:b ħɑ:gɑ:t lil-beɪt وعند اللحظة دي الوضع اتقلب خالص. wi-ʕænd ɑl-læħzˤæ di: ɑl-wæḍʕ itʔælæb χɑ:liş
Adverb of Manner ‘حال’ /ħɑ:l/ **(ADV_MNN)**	It is a noun that describes the circumstances under which an action takes place. كنت جاية تعبانة قوي ومن كتر التعب قعدت سرحانة. kunt gɑ:jjæ tæʕbɑ:næ ʔæwi: wi-min kutr ɑt-tæʕæb ʔæʕædt sarħɑ:næ
Adverb of Degree ‘ ظرف الحال أو درجة الحال’ /zˤærf ɑl-ħɑ:l ʔæw dærægæt ɑl-ħɑ:l/ **(ADV_DGR)**	It is a noun that indicates the intensity of a verb, adjective, or another adverb. It is not found in MSA, but it is added to EGY in the current tag-set schema. أنا حقيقي بحبك جدا. ʔænɑ: ħæʔi:ʔi: bæħibbæk jiddæn كان نفسي أسمع منك كلمة واحدة بس kɑ:n nifsi: ʔæsmæʕ minnæk kilmæ wɑ:ħæ bæs
Pronoun ‘الضمير’ /ɑḍ-ḍæmi:r/ **(PRN)**	It is a word that acts as the subject of a sentence instead of a noun. The pronouns in this category are the disconnected pronouns. The pronouns in this category are: أنا ʔænɑ:, إحنا ʔiħnɑ:, إنت ʔintæ, إنتي ʔinti:, هو huwwæ, هي hijjæ, هما hummɑ:, and إنتو ʔintu:
Relative Pronoun ‘ الاسم الموصول’ /ɑl-ism ɑl-mæwş u:l/ **(PRN_DEM)**	It is a noun that introduces relative clauses. It connects two sentences to give a full meaning. الكلام اللي قولناه كان صح. ɑl-kælɑ:m illi: ʔulnɑ:h kɑ:n şæħħ
Demonstrative Pronoun ‘ اسم الإشارة’ /ism ɑl- ʔiʃɑ:ræ/ **(PRN_REL)**	It is a noun that is used for proximal or distal reference. It is indicated by a tangible sign a person, an animal, a thing, or a place. اللي قولته ده مش عايزة أسمعه تاني والناس دول تنساهم خالص. illi: ʔultu: dæh miʃ ʕæjzæ ʔæsmæʕu: tɑ:ni: win-nɑ:s döl tinsɑ:hum χɑ:liş
Interrogative Pronouns ‘ اسم الاستفهام’ /ism ɑl- istifhɑ:m/ **(PRN_INT)**	It is a noun that introduces a question about something or an action. أنا مش عارفة ده حصل إزاي وإمتى ومين الناس دول أصلا؟ ʔænɑ: miʃ ʕɑ:rfæ dæh ħæşæl ʔizzɑ:j wi-ʔimtæ: wi-mi:n ɑn-nɑ:s döl ʔæşlæn

B Verbs

Verb Sub Classes	Description and Example
Active Verb ‘ الفعل المبني للمعلوم’ /ɑl-fiʕl ɑl-mæbni: lil-mæʕlu:m/ **(VER_ACT:<tense>)**	It indicates the subject of the verb is doing the action. أنا **سمعت** الكلام ده من حد قريب **VER_ACT:PV** ʔænɑ: sæmæʕt ɑl-kælɑ:m dæh min ħædd ʔuræjjib حسن هو اللي **بيعمل** كدة دايما **VER_ACT:IV** ħæsæn huwwæ illi: bi-jiʕmil kidæ dɑ:jmæn بالله عليك **قول** الحق **VER_ACT:RV** bil-læh ʕæli:k ʔu:l ɑl-ħæʔ
Passive Verb ‘ الفعل المبني للمجهول’ /ɑl-fiʕl ɑl-mæbni: lil-mæghu:l/ **(VER_PSV:<tense>)**	It indicates the subject of the verb undergoes the action rather than doing it. It is rarely used in EGY where the pattern /ʔinfæʕæl/ ‘انفعل’ or /ʔitfæʕʕæl/ ‘اتفعل’ is used instead. البلدي **يوكل** **VER_PSV:IV** ɑl-bælædi: **ju:kæl** حاجة كدة على ما **قسم** **VER_PSV:PV** ħɑ:gæ kidæ ʕælæ: mɑ: ʔusim Nevertheless, some passive verbs from MSA are used in some levels of EGY, for example, the passive /qi:læ/ ‘قيل’ ‘be said’.
Non-Conjugated Verb ‘ الفعل غير المتصرف’ /ɑl-fiʕl ɣæjr ɑl-mutaṣærrif/, also known as frozen verb **(VER_FRZ:<tense>)**	It indicates the non-inflected verbs, also known as **frozen verbs**, that are restricted to one tense only. Whereas non-conjugated verbs in MSA may be restricted to perfect, imperfect or imperative tenses, they may be restricted, in EGY, to the perfect or imperative tenses only: والله أنا قولتله الكلام ده قبل كدة لعل و**عسى** يعمل حاجة **VER_FRZ:PV** wɑl-læhi: ʔænɑ: ʔultilu ɑl-kælɑ:m dæh ʔæbl kidæh læʕæl wi-ʕæsæ: jiʕmil ħægæh **هات** اللي معاك ده **VER_FRZ:RV** hɑ:t illi: mæʕɑ:k dæh
Pseudo Verb ‘شبيه الفعل’ /ʃæbi:h ɑl-fiʕl/ **(VER_SUD)**	It is a word that has the same syntactic behavior as verbs in that they take a subject and a predicate, or a sentential complement. **بس** بقى كدة حرام عليك bæs bæʔæ: kidæh ħærɑ:m ʕæleɪk **معلش** يا حبيبي حصل خير mæʕælɪʃʃ jɑ: ħæbi:bi: ħæṣæl χeɪr

C Particles

Particle Sub Classes	Description and Example
Conjunction ‘حرف عطف’ /ħærf ʕæʈf/ **(CNJ)**	A group of particles used to connect elements of equal status in pronunciation or in meaning. مش متأكدة مين قال كدة **يا** محمد **يا** أحمد miʃ mutæʔækkidæ mi:n ʔɑ:l kidæh jɑ: mæħæmmæd jɑ: ʔæħmæd الدخول في علاقة متبعة **أو** مش مريحة حاجة صعبة قوي ɑd-duχu:l fi: ʕælɑ:qæ mutʕibæh ʔæw miʃ muri:ħæh ħɑ:gæ ṣæʕbæ ʔæwi:
Subordinating Conjunction ‘حرف ربط’ /ħærf ræbʈ/ **(CNJ_SUB)**	A group of particles is used to link two clauses in the sentence or two sentences. Some of these articles are still used in EGY: شكلها مجنونة **لكنها** في منتهى العقل ʃæklæhɑ: mægnu:næh lækinnæhɑ: fi: muntæhæ: ɑl-ʕæʔl Others are found but are never used as subordinating conjunction: كان أمله ينجح **بس** للأسف محصلش kɑ:n ʔæmælu jingæħ bæs lil-ʔæsæf mæħæṣæʃ Others are not found in traditional Arabic: **عشان** تخليني ذكرى في دفترك ʕæʃɑ:n tiχælli:ni: zikræ: fi: dæftærik

Particle Sub Classes	Description and Example
Vocative Particle ‘حرف نداء’ /ħærf nidɑ:ʔ/ **(PRT_VOC)**	A group of particles is used to call or alert a person addressed. A noun preceded by a vocative article is called a vocative noun. يا حبيبتي متعمليش في نفسك كدة بالراحة شوية jɑ: ħæbi:bti: mætiʕmili:ʃ fi: næfsik kidæh bir-rɑ:ħæ ʃiwæjjæh
Preposition ‘حرف جر’ /ħærf gærr/ **(PRP)**	A group of particles that is used with a noun, pronoun, or noun phrase to show direction, location, or time or introduce an object. حتلاقيني هناك من بدري ħætlɑ:ʔi:ni: hinɑ:k min badri: موجود على المكتب أو في الدرج mæwgu:d ʕælæ: ɑl-mæktæb ʔæw fi: ɑd-durg
Augment Particle ‘حرف زائد’ /ħærf zɑ:ʔid/ **(PRT_AUG)**	A group of particles that do not affect the meaning if removed from the sentence, but it is added to denote affirmation. ما أنا جبتها لك .. مش جبتها لك mɑ: ʔænɑ: gibtæhɑ: læ-k miʃ gibtæhɑ: læk
Exceptive Particle ‘ حرف استثناء’ /ħærf istiθnɑ:ʔ/ **(PRT_EXC)**	A group of particles used to exclude the following noun from the scope of the words before it. كله إلا كدة والله حرام kulluh ʔillɑ: kidæh wɑl-læhi: ħæɾɑ:m وفي الآخر محدش جه غير نورين wi-fi: ɑl-ʔɑ:xir mæħæddiʃ jæh ɣeɪr nu:ri:n لقيتك أرض متضمش سوى المطاريد læʔeɪtik ʔærḍ mætḍummiʃ siwæ: ɑl-mæṭɑ:ri:d
Emphatic Particle ‘ حرف توكيد’ /ħærf tæwki:d/ **(PRT_EMP)**	A group of particles that used to put emphasis on intention. أما سيدنا النبي ربه كافيه ʔæmmɑ: si:dnɑ: ɑn-nabi: ræbbuh kɑ:fi:h
Futurity Particle ‘ حرف استقبال’ /ħærf istiqbɑ:l/ **(PRT_FUT)**	It is a particle that modifies the verb tense from the present tense to the future. It is not usually used in EGY. قبل أي شيء سوف أسقط الدستور الحالي ʔæbl ʔæjj ʃeɪʔ sæwfæ ʔusqiṭ ɑd-dustu:r ɑl-ħɑ:li:
Negative Particle ‘حرف نفي’ /ħærf næfj/ **(PRT_NEG)**	A group of particles is used to negate the proposition expressed after them, or to deny its affirmation. الموضوع بوخ ومش حلو خالص كدة ɑl-mæwḍu:ʕ kidæ bæwwæx wi-miʃ ħilw xɑ:liʂ الفلم ما كنش حلو خالص ɑl-film mɑ: kænʃ ħilw xɑ:liʂ لأ مش صح læʔ miʃ ʂæħħ أنا لا عايزة أشوفك ولا أسمع صوتك ʔænɑ: lɑ: ʕɑ:jzæ ʔæʃu:fæk wælɑ: ʔæsmæʕ ʂötæk
Explanation Particle ‘ حرف تفسير’ /ħærf tæfsi:r/ **(PRT_XPL)**	A group of particles used to ask to explain the preceding word, phrase or sentence. It is not commonly used in EGY. في يوم عشرة من شوال أي بعد عيد الفطر fi: jöm ʕæʃæræ min ʃæwwɑ:l ʔæj bæʕd ʕi:d ɑl-fiṭr
Interrogative Particle ‘ حرف استفهام’ /ħærf istifhɑ:m/ **(PRT_INT)**	A group of particles is used to elicit understanding, conception, or approval. The noun that follows an interrogative particle is called an interrogative noun. هل ممكن حد فيكم يقولي إحنا وصلنا لهنا إزاي؟ hæl mumkin ħædd fi:kum jiʔulli: ʔiħnɑ: wæṣælnɑ: li-hinɑ: ʔizzɑ:j

Particle Sub Classes	Description and Example
Verb Particle ‘حرف فعل’ /ħærf fiʕl/ **(PRT_VER)**	A group of non-governing particles that precede the perfect or imperfect verbs and do not affect their mood. **وقد** أثبت الإرهاب فشله wæ- qæd ʔæɵbætæ ɑl-ʔirhɑ:b fæʃæluh **قد** يكون الموضوع غريب حبتين qæd jiku:n ɑl-mæwḍu:ʕ ɣæri:b ħæbbitɛn

D Residuals

Residuals	Description and Example
Abbreviation **(ABR)**	It is a shortened form used in place of the whole word or phrase to save space and time, avoid repetition of long words and phrases, or simply to conform to conventional usage. For example, /d/ ‘د’ express the word /duktör/ ‘دكتور’ ‘doctor’.
Emojis **(EMO)**	Any of various small images, symbols, or icons used in texts to express the emotional attitude of the writer, convey information concisely, convey a message playfully, without using words, etc. Examples: 😂 ☹ 👌 👍
Latin Words **(LTN)**	All non-Arabic words are written in other alphabets. ‘good’, ‘responsibility’, and ‘s’Joe’.
Foreign Words **(FRN)**	Non-Arabic words that are written in Arabic alphabets as spoken in another language with no morpholohical changes or adoptations. For example, /weɪr ʔær ju: gö/ ‘وير أر يو جو’.
Numbers **(NUM)**	All alphanumeric numbers.
Punctuation Marks **(PNC)**	They include full stop, comma, colon, semicolon, parentheses, square brackets, quotation mark, dash, question mark … etc.
Interjections **(INJ)**	Words that express the speaker’s reaction to a particular suggestion or sentence. For example, /hhhh/ ‘هههههه’, /ti:t/ ‘تيت’ and /jöh/ ‘يوه’.

E Prefixes

Prefix	Description and Example
Conjunction ‘ حرف عطف’ /ħærf ʕæʧ/ **(CNJ)**	A group of Prefixes that is attached to the beginning of another word to connect elements of equal status in pronunciation or meaning. رحت لحد عنده **وسألته** إيه اللي حصل **فرفض** يقولي أي حاجة ruħt liħædd ʕænduh wi-sæʔltuh ʔeɪh illi: ħæʂæl fæ-ræfæḍ jiʔu:lli: ʔæjj ħɑ:gæh
Definiteness Particle ‘أداة تعريف’ /ʔædɑ:t tæʕri:f’ **(DET)**	It is a definite article that is attached to the beginning of another noun or adjective and makes them definite, rather than indefinite. **الحكاية** وما فيها إن **البنت** دي كانت جميلة جدا وكل **الشارع** بيحبها ɑl-ħikɑ:jæh wi-mɑ: fi:hɑ: ʔinn ɑl-bint di: kɑ:nit gæmi:læh giddæn wi-kull ɑʃ-ʃɑ:riʕ bijħibbæhɑ:
Causative Particle ‘حرف تعليل’ /ħærf tæʕli:l/ **(PRT_CST)**	A group of particles that is attached to the beginning of an imperfect verb to express and confirm the logic of an argument. It is worth mentioning that it is not used in all levels of Arabic in Egypt. لازم نتفق إنه جاء **ليحمينا** لا **ليقهرنا** lɑ:zim nittifiʔ ʔinnuh jɑ:ʔ li-jæħmi:nɑ: lɑ: li-jæqhærnɑ:

Prefix	Description and Example
Preposition ‘ حرف جر’ /ħærf gærr/ **(PRP)**	A group of particles that is attached to the beginning of another noun, or pronoun to show direction, location, or time, or to introduce an object. In traditional Arabic, there are three prepositions that are still used in EGY: /kɑ:f/ ‘ك’, /bɑ:ʔ/ ‘ب’ and /lɑ:m/ ‘ل’. كان **كصديق** مقرب ليها وكانت دايما تحكي **له** كل حاجة **بالتفصيل**. kɑ:n kæ-ṣædi:q muqærræb li:hɑ: wi-kɑ:nt dɑ:jmæn tiħki: lu-h kull ħɑ:gæh biɑt-tæfṣi:l In EGY, the prepositions /bi:/ ‘بي’ and /li:/ ‘لي’ are attached to pronouns: هي كانت **ليها** رموش طويلة hijjæ kɑ:nit li:-hɑ: rumu:ʃ ṭæwi:læh حاول يتصل **بيها** أكتر من مرة بس ما ردتش ħɑ:wil yittiṣil bi:-hɑ: ʔæktær min mærræh The prepositions /fi/ ‘ف’, /ʕæ/ ‘ع’ that are variations of /fi:/ ‘في’ and /ʕælæ:/ ‘على’, respectively, are now used, in EGY, as prefixes. مش فاكر كنت سايبه **عالمكتب** هنا ولا **فالعربية** miʃ fɑ:kir kunt sɑ:jbuh ʕɑl-mæktæb hinɑ: wællɑ: fil-ʕæræbijjæh
Emphatic Particle ‘حرف توكيد’ /ħærf tæwki:d/ **(PRT_EMP)**	A group of particles that is attached to the beginning of a perfect or imperfect verb to put emphasis on intention. أوعى **لينسوك** يا ولدي مصر مين ʔiwʕæ: læ-jnæssu:k jɑ: wælædi: mæṣr mi:n والله لولا تدخل الجيش لحماية الثورة **لكنا** ليبيا جديدة wɑl-læhi: löɬɑ: tædæxxul ɑl-geɪʃ li-ħimɑ:jit ɑs-sæwræh lɑ-kunnɑ: li:bjɑ: gidi:dæh
Futurity Particle ‘حرف استقبال’ /ħærf istiqbɑ:l/ **(PRT_FUT)**	It is a particle that is attached to the beginning of an imperfect verb to represent the future tense. The traditional future particle /sæ/ ‘س’ is rarely used in EGY and the /ħæ/ ‘ح’, /ʕæ/ ‘ع’ and /hæ/ ‘ه’ are used instead. مش **حيكون** أكتر من اللي حصل miʃ ħæ-jkuːn ʔæktær min illi: ħæṣæl وقاللي صدقيني مش **هتندمي** wi-ʔɑ:lli: ṣæddæʔi:ni: miʃ hæ-tindæmi: **عتسأليني** ليه بلملم ف الخلج ʕæ-tisʔæli:ni: leɪh bælæmlim fi: ɑl-*x*ælæg (*Example from Upper Egypt*)
Progressive Particle ‘ حرف للمضارع المستمر’ /ħærf lil-muḍɑ:riʕ ɑl-mustamirr/ **(PRT_PRG)**	A group of particles that is not used in traditional Arabic and is attached to the beginning of an imperfect verb to express the incomplete action or state in progress at a specific time. كان **بيعيد** ويزيد في الكلام كل شوية kɑ:n bi-jʕi:d wi-jzi:d fi: ɑl-kælɑ:m kull ʃiwæjjæh
Jussive-governing Particle ‘حرف جزم’ /ħærf gæzm/ **(PRT_JSV)**	A group of particles that is attached to the beginning of an imperfect verb only to express a required action to do. It is rarely used in EGY. **ولنتابع** الأحداث الجارية بكل حرص wæl-nutɑ:biʕ ɑl-ʔæħdɑ:s ɑl-gɑ:rijjæh bikull ħirṣ

Prefix	Description and Example
Negative Particle ‘حرف نفي’ /ħærf næfj/ **(PRT_NEG)**	A group of particles that is attached to the beginning of another word to negate it or deny its affirmation. This is newly added in EGY. It is not used in traditional Arabic. **مـحدش** في الدنيا يستاهل **ومـفيش** حد الواحد يضحي بنفسه عشانه **مـيستهلوشي** mæ-ħæddiʃ fi: ɑd-dunjɑ: jistɑ:hil wimæ-fi:ʃ ħædd ɑl-wɑ:ħid jiḍæħħi: binæfsuh ʕæʃɑ:nuh mæ-jistæhlu:ʃi:
Vocative Particle ‘حرف نداء وتنبيه’ /ħærf nidɑ:ʔ witænbi:h/ **(PRT_VOC)**	A group of particles that is attached to the beginning of a noun to call or alert a person addressed. **أحبيبتي – أصاحبي – أزميلي - آهو** ʔæ-ħæbi:bti: - ʔæ-ṣɑ:ħbi: - ʔæ-zmi:li: - ʔæ-hu:
Imperative Verb Particles ‘حروف الأمر’ /ħuru:f ɑl-ʔæmr/ **RVPRF_(PGN)**	A group of particles (أ، ن، ي، ت) that are attached to the beginning of the infinitive verb and change it to the present tense without changing its basic form. They are represented in word-form in proclitic-word-form-enclitic representation. اعمل RVPRF_2MS اعملي IVPRF_2FS اعملوا IVPRF_2MP
Imperfect Verb Particles ‘حروف المضارعة’ /ħuru:f ɑl-muḍɑ:riʕæh/ **IVPRF_(PGN)**	A group of particles (أ، ن، ي، ت) that are attached to the beginning of the infinitive verb and change it to the present tense without changing its basic form. They are represented in word-form in proclitic-word-form-enclitic representation. أقول IVPRF_1S يقول IVPRF_3MS يقولوا IVPRF_3MP يقلن (rarely used in EGY) IVPRF_3FP نقول IVPRF_1P تقول IVPRF_2MS or IVPRF_3FS تقولي IVPRF_2FS تقولوا IVPRF_2MP تقلن (rarely used in EGY) IVPRF_2FP

F Suffixes

Suffix	Description and Example
Negative Particle ‘حرف نفي’ /ħærf næfj/ **(PRT_NEG)**	A group of particles that is attached to the end of another word to negate it or deny its affirmation. This is newly added in EGY; it is not used in traditional Arabic. It is always accompanied with the prefix negative particle /mæ/ ‘م’ or the negative particle /mɑ:/ ‘ما’. **محدش** في الدنيا يستاهل **ومفيش** حد الواحد يضحي بنفسه عشانه **ميستهلوشي** mæ-ħæddiʃ fi: ɑd-dunjɑ: jistɑ:hil wimæ-fi:ʃ ħædd ɑl-wɑ:ħid jiḍæħħi: binæfsuh ʕæʃɑ:nuh mæ-jistæhlu:ʃi:
Pronoun ‘ الضمير المتصل’ /ɑḍ-ḍæmi:r ɑl-muttæṣil/ **(PRN)**	A group of pronouns that is attached to the end of a verb and represents its subject or object. It may also be attached to a noun or a preposition (stem or prefix preposition). أنا مش **نبهتكم** من إمبارح للموضوع ده أهو **موبايلها** مقفول ومحدش عارف **لها** طريق والحكاية دي **فيها** إن ʔænɑ: miʃ næbbihtu-kum min ʔimbɑ:riħ lil-mæwḍu:ʕ dæh ʔæhu: mubɑ:jil-hɑ: mæʔfu:l wi mæ-ħæddiʃ ʕɑ:rif læ-hɑ: ṭæri:ʔ wil-ħikɑ:jæ di: fi:-hɑ: ʔinnæ

Suffix	Description and Example
Noun Suffixes **NSUF_(GND)**	A letter or a group of letters (morphemes) that are added to the end of a stem and change the noun gender or number. They are represented in word-form in proclitic-word-form-enclitic representation. ‘أمهات’ /ʔummæhɑ:t/ ‘هات/NSUF_FP’, أبهات /ʔæbbæhɑ:t/ ‘هات/NSUF_MP’ ‘علامات’ /ʕælɑ:mɑ:t/ ‘ات/NSUF_FP’, ‘كتبات’ /kutubɑ:t/ ‘ات/NSUF_MB’ ‘رحمة’ /ræħmæh/ ‘ة/NSUF_FS’, ‘خواجة’ /xæwɑ:gæh/ ‘ة/NSUF_MS’ ‘كتابين’ /kitɑ:beɪn/ ‘ين/NSUF_MD’, ‘ممثلين’ /mumæssili:n/ ‘ين/NSUF_MP’ ‘مشاكل’ /mæʃɑ:kil/ ‘null/NSUF_FB’, ‘أرض’ /ʔærḍ/ ‘null/NSUF_FS’, etc.
Perfect Verb Suffixes **PVSUF_(PGN)**	A letter or a group of letters (morphemes) that are added to the end of a stem and change the perfect verb gender, number or person. They are represented in word-form in proclitic-word-form-enclitic representation. ‘شفت’ /ʃuft/ ‘ت/PVSUF_2MS’ or ‘ت/PVSUF_1S’, ‘شافت’ /ʃɑ:fit/ ‘ت/PVSUF_3FS’ ‘قالوا’ /ʔɑ:lu:/ ‘وا/PVSUF_3MP’, ‘عمل’ /ʕæmæl/ ‘null/PVSUF_2MS’, etc.
Imperfect Verb Suffixes **IVSUF_(PGN)**	A letter or a group of letters (morphemes) that are added to the end of a stem and change the imperfect verb gender, number or person. They are represented in word-form in proclitic-word-form-enclitic representation. ‘يكونوا’ /jiku:nu:/ ‘وا/PVSUF_3MP’, ‘تكتبو’ /tiktibu:/ ‘و/PVSUF_2MP’ ‘يهون’ /jihu:n/ ‘null/PVSUF_3MS’, etc.
Imperative Verb Suffixes **RVPRF_(PGN)**	A letter or a group of letters (morphemes) that are added to the end of a stem and change the imperative verb gender, number or person. They are represented in word-form in proclitic-word-form-enclitic representation. ‘قولي’ /ʔu:li:/ ‘ي/RVSUF_2FS’, ‘ارسم’ /irsim/ ‘null/RVSUF_2MS’, ‘روحوا’ /ru:ħu:/ ‘وا/ RVSUF_2MP’, etc.