A Comprehensive Full-Form Lexicon for Arabic NLP and Speech Technology

Jack Halpern
The CJK Dictionary Institute
34-14, 2-chome, Tohoku, Niiza-shi
Saitama 352-0001
JAPAN
jack@cjki.org

Yannis Haralambous
IMT Atlantique & UMR CNRS 6285 LabSTICC
Technopole Brest-Iroise, CS 83818
29238 Brest Cedex 3
FRANCE
yannis.haralambous@imt-atlantique.fr

Abstract

Natural Language Processing (NLP) applications require morphological data with precise grammatical attributes, while speech technology requires abundant phonemic and phonetic transcriptions. This presents a challenge for Arabic due to its abundant morphological, orthographic, and phonemic variation. Existing systems encounter challenges in processing incomplete and unstructured data from web sources, leading to suboptimal performance in morphological analysis and speech technology. ArabLEX, a comprehensive full-form lexicon for MSA, addresses these issues by providing a foundation for enhancing NLP precision. It comprises over 530 million entries with fully inflected, conjugated, declined, and cliticized forms accompanied by detailed morphological attributes as well as precise phonological transcriptions and orthographic variants. This combines exhaustive listing of forms with detailed descriptions that can significantly mitigate the inherent ambiguity of Arabic. It can serve as a foundation for developing accurate NLP and speech technology applications by providing accurate orthographic variants in both the Arabic script and in phonemic transcriptions.

1 Introduction

1.1 What is a Full-Form Lexicon

According to Crystal's *Dictionary of Linguistics* (Crystal, 2008), a word is a "unit of expression which has universal intuitive recognition by native speakers." Although this does not provide an objective criterion for "wordhood," words are an important notion in NLP. Ever since the first dictionary in history, the *Sumerian Lexicon* (2300 BCE), lexicographers have worked on the collection of "words." Traditionally, the headwords of dictionaries have been canonical forms (lemmata). A rare dictionary format is the *full-form lexicon.* It explicitly includes all word forms of a language, i.e., fully inflected, conjugated, declined, or cliticized ("inflected" for short) members of a lexeme class, rather than just the lemmata. For example, the English lexemes eat and boy have the members *eat*, *eats*, *eating*, *eaten*, *ate* and *boy*, *boys*, *boy's*, *boys'* respectively. For highly inflected languages like Arabic, the abundance of combinatorics (stem, affixes, clitics) can result in full-form lexicons with hundreds of millions of entries.

Historically, Machine Translation (MT) and other NLP applications relied on rules or statistical models. In recent years, utilizing neural machine translation (NMT) and large language models (LLM) has become the norm. These tools rely on efficient disambiguation. Despite their remarkable achievements, challenges remain in Arabic NMT, such as the handling of proper nouns and multiword expressions (MWE) (Halpern, 2019) and overcoming the lack of bilingual training corpora.

1.2 The Case of Arabic

Arabic is special in its morphology. On the one hand, Modern Standard Arabic (MSA), used in the media, government, and education, is the official language of 380 million people, but (practically) no one's mother tongue (Haugen, 1972; Mejdell, 2014). A kind of hierarchical polyglossia is the norm. On the other hand, the morphology of Arabic is based on roots and patterns (templatic morphology) (Ryding, 2005), so we are not just dealing with stems and affixes as in Roman languages but with tri- or quadriliteral consonantal roots with infixes, prefixes, suffixes and circumfixes. This morphological generative principle is omnipresent and

even applies to loanwords (Gadelli, 2015). It can thus be considered to be an innate property of Arabic. Therefore, a full-form lexicon should cover all MSA root + pattern combinations (so that all grammatical word forms are available to the user), which is necessary for both speech recognition and written text.

1.3 Previous Work

Several Arabic modeling tools have been developed for morphological analysis, tokenization, generation of inflected and conjugated forms, POS tagging, and disambiguation. We refer to such tasks as analysis and generation, and to such tools as morphological engines. Popular tools include AlKhalil (Boudchiche *et al.*, 2017), MADA (Habash, Rambow, and Roth, 2009), BAMA (Buckwalter, 2002), PATB (Penn Arabic Treebank) (Maamouri *et al.*, 2004), FARASA (Abdelali *et al.*, 2016), MADAMIRA (Pasha *et al.*, 2014), and Elixir_FM (Smrž, 2007). A more recent, highly ambitious tool is CALIMA Star (Taji *et al.*, 2018).

Despite the high performance of these tools (Taji *et al.*, 2018), they have shortcomings, such as inconsistency, ignoring lexical rationality, and lacking phonological attributes. Naturally, the processing performed by morphological engines is supported by lexical databases, such as tables for stems, clitics, and affixes (Halpern, 2018). Still, the goal of these tools is to perform computational tasks such as tokenization and disambiguation rather than serving as comprehensive lexicons for enumerating all possible ambiguous sequences. A notable outlier worth mentioning is the Arabic full-form lexicon and Finite State Transducer (FST) project by Soudi and Eisele (2004).

1.4 Introducing ArabLEX

Unlike morphological engines, ArabLEX is a stand-alone lexical database that can be integrated with such engines. It does not perform computational tasks itself. Its goal is to act as a comprehensive database to support morphological engines and NLP tools. In theory, an engine can query the lexicon as an external module via function call or API, but ideally, it should be integrated directly. If a morphological engine is likened to the engine of a car, then a full-form lexicon like ArabLEX is the fuel – it drives the engine, not the car itself (e.g., a text-to-speech (TTS) application that relies on a pronunciation dictionary).

ArabLEX is a *full-form lexicon* aiming to be as comprehensive as possible, though some word classes, such as periphrastic elatives, have not yet been included. In the first phase (May 2024) ArabLEX contains about 530 million entries for content words (nouns, adjectives, and verbs) in the domains of general vocabulary and (for the first time) fully inflected and cliticized proper nouns for both Arab and non-Arab personal names and place names. It provides exhaustive coverage of all inflected, declined, conjugated and cliticized forms and includes a rich set of grammatical, morphological, phonological, and orthographic attributes, as shown in detail by The CJK Dictionary Institute (2020). This makes it suitable for NLP applications such as machine translation, named entity recognition, and morphological analysis and generation. For example, the verb *katabta* is one of 7,251 possible forms of *kataba*. It has tags such as 2SM, meaning second person masculine singular. Special emphasis is placed on speech technology by providing such attributes as accurate phonemic transcriptions as well as full diacritization.

Note that the phonemic[1] transcriptions in this paper are italicized and given in the CARS system (Halpern, 2009), designed by our institute for pedagogical and speech applications. Transliterations are given in the Buckwalter transliteration system (Buckwalter, 2002) and enclosed in forward-slashes. Note also that ArabLEX is undergoing maintenance and expansion, and it is expected to exceed one billion entries, making it—to our knowledge—the most comprehensive Arabic computational lexicon ever created.

2 Levels of Ambiguity in Arabic

2.1 Morphological / Lexical Ambiguity

In templatic morphology, inflection is performed by changing the vowel + consonant patterns by affixation and cliticization. Not only can words be inflected, declined, and conjugated ("inflected" for short), but they can also take many clitics. For example, adding the proclitics *wa* 'and', *li* 'to', and the enclitic *ātíhimạ* to the stem *kắtib* 'writer' yields the complex form *walikạtibātíhimạ* (وَلِكَاتِبَاتِهِمَا) 'and to their (dual) female writers'. This results in

[1] Technically, CARS is a morpho-phonemic transcription system, as it encodes information such as vowel neutralization.

a very large number of word forms. For example, the full paradigms for كَاتِبٌ *kātibun* 'writer' and كَتَبَ *kataba* 'write' reach about 5,660 and 6,900 forms, respectively.

The difference between morphological and lexical ambiguity is analogous to the difference between inflection and derivation in Western languages: when a word is inflected, the forms we obtain represent the same lexeme; when it is derived, we move to a different lexeme. This happens also in Arabic, e.g., the graphemic sequence كتب may denote كَتَبَ 'I wrote,' or كُتُبٌ 'books.' The lexeme of the former is the verb 'to write,' and the lexeme of the latter is the noun 'book.'

Distinguishing between morphological and lexical ambiguity is computationally relevant because the latter implies multiple POS tags and, therefore, also potentially multiple syntax trees.

2.2 Recognition of Arabic Patterns

Conventional wisdom has it that Arabic is ambiguous "due to the non-representation of short vowels." In fact, a whole gamut of factors contributes to ambiguity (Halpern, 2002), including (1) the absence of short vowels (e.g., كاتب represents the seven word forms *kātib*, *kātibun*, *kātibin*, *kātaba*, *kātibi*, *kātiba*, *kātibu*), (2) representation of long *ā* by ا as in سوريا or by آ as in آسيا, but some bare alifs representing *tanwiin* rather than long *ā*, as in شكرا *shukran*, (3) *ʾalif alfaaSila* (otiose *alif*) (Ryding, 2005), orthographic conventions not being pronounced (e.g., كتبوا being realized as *katabu*[2]), (4) the omission of *shadda* indicating consonant gemination, e.g., محمد (diacriticized مُحَمَّد), which provides no clues that the /m/ is geminated, and (5) vowel neutralization sometimes being lexically determined and thus unpredictable from the orthography, e.g., في القاهرة 'in Cairo', the preposition /fyi/ is pronounced *fi*, not *fii*.

Examples (1)–(4) given above are cases of graphemically under-represented patterns. Indeed, patterns may contain short vowels or consonants/long vowels that are written but must be recognized as being part of a pattern.

2.3 Orthographic Disambiguation

A central issue in Arabic NLP, especially speech technology, is identifying which word form an ambiguous string like كاتباتك represents. This string can represent any of six-word forms, each with a different meaning, a different pronunciation and potentially a different POS tag and/or syntactic function.

The process of identifying the correct form is referred to as orthographic disambiguation (Halpern, 2008). The rich set of grammatical and morphological attributes in ArabLEX can help language models to correctly disambiguate such forms.

2.4 Word Stress and Vowel Neutralization

Prosody (word stress) and vowel neutralization play a critical role in ensuring that synthesized speech sounds natural. نا *naa*, for example, is written as a long vowel in أنا but is shortened to *na*. This complex issue is described in detail in Halpern's (2009) paper on Arabic stress.

The morpho-phonemic and phonetic transcriptions in ArabLEX explicitly indicate precise word stress and vowel neutralization for each entry. For example, in the IPA [wĕlikĕːˈtibikumɛ(ˑ)], the stressed syllable is indicated by (ˈ) (U+0C28), while (ˑ) (U+02D1) indicates that the final ɛ is a neutralized vowel of optional half-length.

3 Enhancing Speech Technology

3.1 Arabic Speech Technology

Though advances in neural networks have dramatically improved the quality of speech technology, in a 2020 survey, we compared the TTS systems provided by leading IT companies, showing that Arabic significantly lags behind other major languages (Halpern, 2020). ArabLEX addresses these shortcomings by serving as a comprehensive pronunciation dictionary that enhances the quality of both TTS and automatic speech recognition (ASR). It includes an NLP-oriented morpho-phonemic transcription called CARS (Halpern, 2009) and two phonetic transcriptions: SAMPA (Wells, 1997) and IPA (International Phonetic Association, 1999), to support the training of TTS and ASR models. For example, the entry وَلِكَاتِبَاتِهِمَا is transcribed as *wal-ikatibātihimā*, an accurate phonemic representation

[2] Pronouncing وا as *wa* is a grave mistake committed by at least one of the major engines.

to which we have added morphological information. It consists of the stem *kā́tib* 'writer' and the proclitics *wa* 'and' and *li* 'to' and the enclitic *ā́tíhimą* 'their'. The *ą* indicates two occurrences of *ā* that have been neutralized to short *ą*, indicated by the underline.

3.2 TTS Accuracy

Due to the extreme orthographic ambiguity of Arabic, even major IT players struggle to synthesize speech accurately. The CJKI survey (Halpern, 2020) revealed that it is not unusual for over 50%, and even 80%, of the words in a sentence, especially cliticized words, to be mispronounced. Errors are evaluated within the context of a sentence. A pronunciation is considered erroneous if it includes mistakes such as incorrect case endings (e.g., pronouncing الكاتب as *lkā́tibi* when it should be *lkā́tibu),* omitted shaddas (such as pronouncing عدد as *ɛádada* when it should be *ɛáddada* 'to enumerate'), or other pronunciation errors that can be unambiguously identified. In Table 1, pronunciation errors are marked by an asterisk**.**

Unvocalized	Vocalized	Google (13%)	iOS (31%)	Bing (25%)	CJKI
عدد	عَدَّدَ	*ɛádadu	*ɛádadɑ	*ɛádadɑ	ɛáddadɑ
الكاتب	ٱلْكَاتِبُ	*lkā́tibi	lkā́tibu	lkā́tibu	lkā́tibu
ما	مَا	*mą	*mą	*mą	*mą
الحكام	ٱلْحُكَّامَ	*lḥukkā́mi	*lḥukkā́mi	*lḥukkā́mi	lḥukkā́ma

Now let us look at the errors in context for a composed text. The original sentence

عدد الكاتب ما قال إن هؤلاء الحكام يفعلونه في الخارج مثل الهجمات الإلكترونية ومطاردة المعارضين اللاجئين في العواصم الغربية.

was mispronounced by Google TTS as

**ɛádadu [ɛáddada] lkā́tibu mą qā́la ʾínna haʾulā́ʾi *lḥukkā́mi [lḥukkā́ma] yafɛalūnahu fī lkhā́riji *míthli [míthla] lhajamā́ti lʾilikturųníyyati wamuṭārádati lmuɛariḍīna llaji'īna fī lɛawā́ṣimi lgharbíyyati.*

Asterisks mark incorrectly pronounced words. The correct pronunciations are given in brackets.

A more recent test (December 2021) on تَصْحَبُوا with Google produced *taṣḥábųwa,* instead of *táṣḥabų.* Not only is word stress incorrect, but the final وا (*wa* + otiose *alif*) (Ryding, 2005), which must be silent, is pronounced (a major error). Table 1 reveals that the error rate for the major TTS engines is notable, highlighting a substantial need for improvement. The word error rates (WER) in composed texts ranged from 13% to 25%, whereas in web-extracted texts, they ranged from 70% to 90%.

Since the formal tests were conducted, we continued to conduct informal tests of TTS accuracy, such as with Google Translate, and did not observe any significant improvement in accuracy.

3.3 Enhancing TTS accuracy

The CJKI PATTS (Palestinian Arabic Text to Speech) white paper (The CJK Dictionary Institute, 2023) presents samples of an early-development proprietary TTS solution for Palestinian Arabic. It utilizes phonetic data (specifically IPA) to ensure that the underlying TTS system generates accurate realizations. By presenting accurate phonetic data to a TTS system that supports supplementing such data, PATTS is able to generate accurate phonetic realizations without having to fine-tune the underlying TTS system to the Palestinian Dialect. One such system is Amazon AWS Polly, which supports the "phoneme" SSML tag for its Arabic TTS voices, which allows the user to specify how a word should be pronounced using a subset of X-SAMPA or IPA. (Amazon Web Services, 2023)

3.4 ASR Accuracy

For TTS, it is necessary to generate one accurate pronunciation, but ASR systems must recognize alternative pronunciations, including informal ones. For example, the standard pronunciations of كاتبون 'writers' and أكتب 'I write' are *kątibū́na* and *áktubu,* but the less formal variants *kątibū́n* and *áktub* are widespread and possibly even more common.

Such alternatives include pausal forms and final vowel elision. The former refers to sentence-final forms causing final vowels to be elided in Classical Arabic, while the latter is the elision of certain final vowels in both medial and final forms, common in spoken MSA. For example, رَجَعْتُ إِلَى ٱلْبَيتِ 'I returned home', pronounced *rajáɛtu ʾílą‿lbáyti,* in pausal form becomes *rajáɛtu ʾílą‿lbayt* and in spoken MSA becomes *rajáɛt ʾílą‿lbayt.* Note how the final *ti* and *tu* are truncated to *t.*

The above alternatives are for standard MSA. There are also regional allophones. For example, /j/ in words such as *jamal* 'camel' is pronounced [g]

in Egypt, [dʒ] in the Gulf region, and [ʒ] in the Levant. These are regional variants of MSA. ArabLEX not only includes the IPA for the standard MSA, namely [dʒ] for /j/, but also the regional allophones [ʒ] and [g]. It aims to include transcriptions of common non-standard regional allophones.

In this context, applying a phonetic alphabet like IPA or X-SAMPA becomes particularly relevant, as utilizing phonetics in ASR systems has proven beneficial (Feng et al. 2023).

4 Machine Translation

Although NMT has dramatically improved translation quality, it has some shortcomings, as Philipp Koehn (2020) and Halpern (2018) pointed out. Some issues in Arabic are (1) the high orthographic ambiguity, (2) the morphological complexity (forms like ولكاتباتهما are difficult to analyze), (3) the recognition of named entities (often cliticized), and (4) a large number of word forms for nouns and verbs.

ArabLEX offers comprehensive coverage of inflected and cliticized forms and can be used to supplement existing corpora or as a pseudo-corpus for language model training, enhancing the accuracy of morphological, syntactic, and semantic analysis. Additionally, the proper noun modules of ArabLEX – DAN (Database of Arabic Names), DAF (Database of Arabic Foreign Names), and DAP (Database of Arabic Places) – are bilingual and romanized, serving as a bilingual dictionary.

5 ArabLEX in Action

5.1 Scope and Coverage

The first release of ArabLEX in 2021 covered about 530 million entries for general vocabulary and proper nouns. ArabLEX consists of the following four main modules: DAG (Arabic General Vocabulary, 83M entries), DAN (Arabic Names, 218M entries), DAF (Arabic Foreign Names, 226M entries) and DAP (Arabic Place Names, 6M entries). ArabLEX has 30 data fields with detailed grammatical, phonological, morphological, and orthographic attributes (Halpern, 2020).

It can be argued that generating entries by rules and templates can result in a large number of non-existing or erroneous forms. We have taken extreme care to ensure that only grammatically and, as far as possible, semantically valid forms are included. Though currently some forms may not have been observed to exist, they are indeed valid and could occur in the future. For most applications, the negative effects are negligible compared to a lack of data (Koperniak, 2017).

5.2 ArabLEX Compared to Other Resources

The comprehensiveness of a dataset holds significant importance for NLP applications. This is particularly pronounced in morphological engines, (Attia et al., 2011).

Previous efforts to compile extensive Arabic lexicographical or lexical databases have yielded datasets containing around 200,000 unique lemma entries. These datasets tend to lack a diverse set of attributes (Attia *et al.*, 2011; Alshargi *et al.*, 2019). In contrast, the CALIMA dataset for Egyptian Arabic comprises 36,000 distinct lemmata and contains 40 fields with attributes such as morphology, gender, and root (AlShuhayeb, 2023; Habash *et al.*, 2012). Detailed datasets like this typically contain entries in the range of 30,000 headwords (Alshargi *et al.*, 2019).

ArabLEX, on the other hand, covers a combined 375.335 unique lemmata, including a large number of named entities, while exceeding the level of detail and versatility of its counterparts. Especially by offering phonetic (IPA, XSAMPA) and morphophonemic (CARS) transcriptions and fully diacritcized Arabic, ArabLEX fills a gap in current lexical resources.

Another key difference is the total number of entries accessible for explicit analysis; that is, entries that are pre-generated as opposed to on-the-fly. The Calima dataset contains approximately 48 million entries that can be examined when all its lemmata and affixes are exhaustively generated (AlShuhayeb, 2023). By contrast, ArabLEX consists of 530 million entries pre-compiled in TSV format immediately accessible for use and analysis.

5.3 Comparison with CALIMA Star

ArabLEX's model of Arabic morphology is more refined than those of other systems. This results in high recall by covering almost all word forms. Multiple layers of sanity-checking ensure high precision and grammatical validity of each form. To illustrate this, we compared some features of ArabLEX and CALIMA Star ("Calima" below), the most advanced morphological engine, using the

affirmative of the verb كَتَبَ ‘to write’. The results are based on the Calima generator web interface.

(1) The coverage of inflected and cliticized forms differs dramatically. Many conjugated forms are missing in Calima, which also generates some invalid forms. The table below shows the number of forms for كَتَبَ.

Item	CALIMA Star	ArabLEX
Total forms	2,448	5,886
Uncliticized	104	124
Cliticized	2,344	5,762

Table 2: Coverage CALIMA Star vs. ArabLEX.

(2) The cliticized forms كَتَبْتُنَا ,كَتَبْتِنِي and كَتَبْتَكَ are not given by Calima, whereas some forms it provides, like لَايَكْتُبُ, are grammatically invalid. The number of cliticized forms provided by ArabLEX (both proclitics and enclitics) for كَتَبَ exceeds that of Calima by 146%.

(3) The results of a preliminary investigation of proclitic coverage by Calima (expanded on below) shows that Calima does not support the proclitic />a/ (أَ), even if selected from the menu. ArabLEX provides more clitic combinations: 39 proclitic combinations and over 2000 (to our knowledge double that of Calima) proclitic-enclitic combinations, which were carefully vetted to ensure their validity. For example, the singleton proclitic sequence />awabi{lo/ is a valid combination for nouns, but />awaka{lo/ is not, while any proclitic in />a, wa, fa, >awa, >afa/ can combine with any enclitic in /N, FA, FY/ for singular nouns.

(4) Whereas ArabLEX takes great care to include only grammatically valid forms, Calima seems to generate agrammatical forms such as سَأَكْتُبَ and سَتَكْتُبَ, or invalid forms such as لَاأَكْتُبُ instead of لَا أَكْتُبُ (omitting the space after لَا).

(5) The verb conjugation paradigm is missing important forms. For example, Calima does not return the active participle كَاتِبٌ, nor the passive participle مَكْتُوبٌ for the verb lemma كَتَبَ.

(6) The imperative forms اُكْتُبِي ,اُكْتُبْ, etc. are not generated even when explicitly requested via the user interface, which is a serious shortcoming.

In conclusion, the ArabLEX morphological model of the verb كَتَبَ is significantly finer than that of Calima. Whereas the former always provides grammatically valid forms, the latter sometimes generates agrammatical ones.

5.4 Real-World Applications

Amazon has acknowledged the significant contribution of ArabLEX to its advanced Arabic speech technology for Alexa. It’s comprehensiveness and the in-depth morphological and phonological data has helped Amazon reduce the error rates for both recognition and generation; that is, to recognize Arabic queries, including place and personal names, as well as generate answers with greater precision (The CJK Dictionary Institute, 2022).

5.5 Grammatical Attributes

The grammatical attributes of ArabLEX are useful for morphological analysis, orthographic disambiguation, POS tagging, semantic analysis, and more. These include codes for gender, number, case endings and person, as well as the stem, definiteness, lexical rationality, and the lemma.

Data field	Value
Full-form	وَلِكَاتِبِكُمَا
Lemma	كَاتِبٌ
Stem	كَاتِب
Gender	C (common)
Case	GEN (genitive)
Number	D (dual)
Person	2 (second)
Definiteness	D (definite)
Root	ك-ت-ب

Table 3: Grammatical attributes.

5.6 Phonological Attributes

The phonemic and phonetic transcriptions are useful for improving speech technology, both TTS and ASR (Tahon *et al.*, 2016; Feng *et al.*, 2023). These include precise, fully diacriticized Arabic with accurate phonemic and phonetic transcriptions as well as word stress and vowel neutralization. The main phonological attributes are shown in Table 4.

Data field	Value
Vocalized	مُحَمَّدٌ
Phonemic	*muhammadun*
Phonetic	[muˈħɛmmɛdun]
X-SAMPA	mu"X\E_"mmE_"dun
Transliterated	muham~adN

Table 4: Phonological attributes for محمد.

5.7 Morphological/Orthographic Attributes

The morphological attributes include all inflected, conjugated, declined, and cliticized word forms, such as plurals, duals, feminine, case endings, conjugated forms, as well as proclitics, enclitics, stems, and roots. They are useful for morphological analysis, semantic analysis, lemmatization, decliticization, deaffixation, verb conjugation, and dictionary lookup. Operations such as decliticization, deaffixation and tokenization (Carbonell et al., 2006) are easy to perform since clitics are given explicitly in their own fields (Enclitic, Proclitic**,** and Stem below). The main morphological attributes are shown in Table 5.

Data Field	Value	Transcription
Full-form	ولكاتبكما	*walikātibikumạ*
Lemma	كاتب	*kắtibun*
Stem	كاتب	*kắtib*
Proclitic	ول	*wali*
Enclitic	كما	*(i)kúmạ*
Root	ك-ت-ب	*k-t-b*

Table 5: Morphological attributes.

Orthographic attributes are useful for orthographic disambiguation, which is necessary for word and entity recognition, TTS, morphological analysis, word/entity extraction, normalization, and dictionary lookup. These include orthographic variants such as pausal and elided forms and even common typographical oddities. Here is an example of typical orthographic variants for the name Alexandra: ألكسندرة, الكسندرة, ألكسندره, الكسندره, ألكسندرا, الكسندرا. As shown above, ه and ة are sometimes interchangeable in names. Orthographic variants also include allographs, for example the use of ى (*alif maqsuura*) as an alternative for ي (*yaa*) in Egypt, and the use of پ instead of ب for [p] in some regions.

5.8 Named Entity Recognition

The DAN module of ArabLEX covers about 100,000 vocalized personal names and their 6.5 million romanized variants. DAN is widely deployed in both security and NLP processing tools for NER and MT. Similarly, the DAF and DAP modules consist of about 240,000 names for places and non-Arab personal names. These modules account for about 450 million fully inflected and cliticized entries in ArabLEX (Halpern, 2009).

5.9 Accessing ArabLEX

ArabLEX is specifically targeted at researchers and software developers needing rich morphological and phonological resources. It is available through The CJK Dictionary Institute and the European Language Resource Association, a non-profit repository of language resources (European Language Resource Association, 2022).

6 Compilation Methods

6.1 Quality Control

The ArabLEX team, comprising professional editors, translators, computational linguists, and university instructors, has conducted extensive research to ensure maximum accuracy and comprehensive coverage of all word forms and their variants. Many programs were developed for data validation and proofreading to ensure accuracy and consistency, such as programs for automatic error detection and correction and data validation. The following outlines one of the data validation processes our team employs to refine our vocalization validation module (VBW_INTEG) to ensure the accuracy of fully vocalized Arabic and phonemic transcriptions, critical for speech technology:

(1) A program validates the fact that inflections are correctly vocalized based on strictly defined rules such as *hamza* rules, presence of short vowels and many more. (2) The program then attempts to rectify the errors it encounters autonomously. (3) Errors that the program cannot rectify are presented to our proofreaders, who manually classify, analyze, and rectify them. (4) Based on the feedback of our proofreaders, the validation rules are then either adjusted or our database of exceptions is expanded. (5) The process is then repeated.

This iterative process has been applied over the course of many years, resulting in a system with a comprehensive set of rules and exceptions.

To illustrate, when validating one of our Arabic dictionaries using the same program, we identified entries such as /diywAnN/ (ديوانٌ) and /$>owN/ (شأْوٌ). Subsequently, the spellings were automatically corrected to /diywaAnN/ (ديوَانٌ) and /$a>owN/ (شَأْوٌ). Our proofreaders then reviewed these modifications and confirmed them as correct.

Note that this process has refined our rule base to be highly sophisticated. True exceptions are rather rare, normally one-off isolated instances. If a trend or pattern is found in the exceptions, they are analyzed and codified as rules in the error detection program so they will no longer be considered errors. The accuracy of phonemic transcriptions is likewise ensured, as it uses the fully vocalized Arabic generated by this process and undergoes a similar validation process.

6.2 Inflection, Conjugation, Cliticization

Generating inflected forms involves many complex steps, including sanity checking and human proofreading. Nouns and adjectives are declined/inflected for feminine, dual, and plural forms. For example, for /bayotN/ 'house,' we derive /bayotaAni/, /buyuwtN/, and /buyuwtaAtN/. As for conjugation, the verb paradigms from the CJKI Arabic Verb Conjugator (CAVE) (The CJK Dictionary Institute, 2011) are used to acquire the verb conjugations for each subject pronoun for each tense. CAVE has 180 categories and fully explicit conjugated paradigms (generated by hand-vetted precise rules and exceptions) for each category. For example, for /kataba/ 'he wrote' we get /yakotubu/ (third person masculine singular imperfect), /Aukotubo/ (second person masculine singular imperative), etc. To encliticize, the correct enclitic template is selected based on the ending of the inflected form. For example, the noun /|xirapu/ 'the hereafter' ends in /pu/, so the template in Table 6 is selected. Enclitics are then added to correspond to each case and subject pronoun. For /|xirapu/, we generate such forms as /|xiratiy/, /|xiratuka/ and /|xiratuki/. To procliticize, the appropriate proclitics are elected from the template. For example, for /bayotN/ 'house', the enclitic is /-N/ (tanwiin), so we refer to the appropriate row (row 2) in Table 7 and generate />abayotN/, /wabayotN/, etc.

Note that the clitics are not merely blindly concatenated to the base form – there are over 2,000 valid orthographic, grammatical, and semantic combinations of clitics that are defined by our human-vetted constraint-defining tables, as shown in Table 7, and several thousand that are invalid.

Per	Case	Enclitic	Rule
000	NOM	u	
1SC	NOM	iy	-p → -t
2SM	NOM	uka	-p → -t
2SF	NOM	uki	-p → -t

Table 6: Template for nouns that end in /p/ (ة).

Proclitic	Enclitic	Gen	Num
0,>a,wa,fa,>awa,>afa,Aalo,...	a,u	M	S
0,>a,wa,fa,>awa,>afa	N,FA,FY	M	S
0,>a,wa,**fa,**>awa,>afa	uhaA,uhu,uhumaA, uhumo, uhun~a,uka, uki,ukumaA,...	M	S

Table 7: Possible combinations of clitics.

7 Future Work

We will continue expanding ArabLEX by adding new entries and data fields, including technical terms, and named entities, as well as more phonological attributes, orthographic variants, alternative pronunciations, and additional word classes (POS). Especially noteworthy are new headwords that consist of multiword expressions (Halpern, 2019) (inflections or conjugations consisting of space-delimited components), such as periphrastic elatives (أَكْثَرُ إيلَام 'more painful'), negative elatives (with أَقَلُّ or أَخَفُّ), inflected numerical expressions, phrasal verbs, compound tenses, verb negation, and more. The addition of proclitics, enclitics and inflections lead to ArabLEX exceeding 500 million records (15 billion data points). It is expected to reach about one billion records in the near future.

In parallel to ArabLEX, we have been developing a series of full-form lexicons for the major Arabic dialects, called DiaLEX, based on the same methodology used for ArabLEX. Since there is no official orthography for the dialects, we conducted thorough research on various dialects by analyzing corpora and dictionaries and by collaborating with native-speaking experts. As a result, we have identified the most common conventions for each dialect, which made the creation of DiaLEX possible. DiaLEX currently (May 2024) covers the major Arabic dialects Egyptian, Emirati and Hijazi. The initial release of the first three has been completed, covering about 150 million entries, and the development of a Palestinian full-form lexicon (PA_LEX) is now in progress (May 2024).

ArabLEX, in combination with DiaLEX, can serve as a holistic resource for the development of NLP applications for MSA and its dialects.

8 Bibliographical References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California, June. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3003

Algihab, W., Alawwad, N., Aldawish, A., and Al-Humoud, S. (2019). Arabic Speech Recognition with Deep Learning: A Review. In G. Meiselwitz (Ed.) *Social Computing and Social Media. Design, Human Behavior and Analytics*, Springer International Publishing (pages15-31). https://doi.org/10.1007/978-3-030-21902-4_2

Alshargi, F., Dibas, S., Alkhereyf, S., Faraj, R., Abdulkareem, B., Yagi, S., Kacha, O., Habash, N., & Rambow, O. (2019). Morphologically Annotated Corpora for Seven Arabic Dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (pp. 137-147). https://doi.org/10.18653/v1/W19-4615.

AlShuhayeb, H., Minaei-Bidgoli, B., Shenassa, M. E., Hossayni, S (2023). Noor-Ghateh: A Benchmark Dataset for Evaluating Arabic Word Segmenters in Hadith Domain. arXiv preprint arXiv:2307.09630. https://doi.org/10.48550/arXiv.2307.09630

Amazon Web Services, Inc. (2023). *Amazon Polly Developer Guide*. https://docs.aws.amazon.com/polly/

Attia, M., Pecina, P., Toral, A., Tounsi, L., & Genabith, J. (2011). An Open-Source Finite State Morphological Transducer for Modern Standard Arabic. In *Proceedings of the Fourth International Conference on Arabic Language Resources and Tools (MEDAR)*, pp. 125-133. https://aclanthology.org/W11-4417

Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., Lakhouaja, A., and Boudlal, A. (2017). "AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer," *Journal of King Saud University – Computer and In*formation Sciences 29 (2), 141–146. https://doi.org/10.1016/j.jksuci.2016.05.002

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2002L49. https://doi.org/10.35111/7vzm-mb15

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., and Frei, J. (2006). Context-Based Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 19-28, Cambridge, Massachusetts, USA, August. Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.3/

CJK Dictionary Institute, The (2011). The CJKI Arabic Verb Conjugator. Downloaded from <http://cjki.org/arabic/cave/cavehelp.htm> on 13 January 2022.

CJK Dictionary Institute, The (2020). ArabLEX Technical Specifications. Downloaded from <https://www.cjk.org/wp-content/uploads/Arablex_specs.pdf> on 13 January 2022.

CJK Dictionary Institute, The (2020). Enhancing Arabic Speech Technology with comprehensive Arabic training lexicon. Downloaded from <https://www.cjk.org/wp-content/uploads/TTS_Report.pdf> on 13 January 2022.

CJK Dictionary Institute, The (2022). ArabLEX Arabic Full-Form Lexicon. Retrieved from <https://www.cjk.org/data/arabic/nlp/arablex-arabic-full-form-lexicon/>

CJK Dictionary Institute, The (2023). *Palestinian Arabic Text-to-Speech system (PATTS)*. Retrieved from <https://www.cjk.org/wp-content/uploads/patts.html>

Crystal, D. 2008. *A Dictionary of Linguistics and Phonetics*. New York: Wiley-Blackwell. ISBN 9781405152969. https://doi.org/10.1002/9781444302776.

Diab, N. (2021). Out of the BLEU: An Error Analysis of Statistical and Neural Machine Translation of WikiHow Articles from English into Arabic. *CDELT Occasional Papers in the Development of English Education*. 75. 181-211. https://doi.org/10.21608/opde.2021.208437.

European Language Resources Association (ELRA). (2022). *The Multilingual European Language Corpus* [Datasets: ELRA-L0131, ELRA-M0105, ELRA-M0106 and ELRA-M0107]. ELRA. https://catalog.elra.info

Feng, S., Tu, M., Xia, R., Huang, C., & Wang, Y. (2023). Language-universal phonetic encoder for low-resource speech recognition. arXiv. https://arxiv.org/abs/2305.11576

Gadelli, N. 2015. The morphological integration of loanwords into Modern Standard Arabic: Towards a morphological categorization of loanwords. Bachelor thesis in General Linguistics, Lund University, Sweden. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=5053170&fileOId=5053172

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS

150 tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium. https://api.semanticscholar.org/CorpusID:10124099

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*, pages 1–9, Montréal, Canada. Association for Computational Linguistics. https://aclanthology.org/W12-2301

Halpern, J. (2002). The Challenges and Pitfalls of Arabic Romanization and Arabization. Downloaded from <https://www.cjki.org/arabic/arannana.pdf> on 13 January 2022.

Halpern, J. (2008). Exploiting Lexical Resources for Disambiguating CJK and Arabic Orthographic Variants. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/109_paper.pdf

Halpern, J. (2009). Word stress and vowel neutralization in modern standard Arabic. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium. http://www.elda.org/medar-conference/pdf/16.pdf

Halpern, J. (2009). CJKI Arabic Romanization System (CARS). Downloaded from <https://www.cjki.org/cjk/arabic/cars/cars_paper.pdf> on 13 January 2022.

Halpern, J. (2009). Lexicon-Driven Approach to the Recognition of Arabic Named Entities. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium. https://api.semanticscholar.org/CorpusID:43949907

Halpern, J. (2016). Compilation Techniques for Pedagogically Effective Bilingual Learners' Dictionaries. *International Journal of Lexicography*, Volume 29(3): 323–338. https://doi.org/10.1093/ijl/ecw023

Halpern, J. (2018). Very large-scale lexical resources to enhance Chinese and Japanese machine translation. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA). https://aclanthology.org/L18-1137

Halpern, J. (2019). Lexicographic Criteria for Selecting Multiword Units for MT Lexicons. Downloaded from <https://www.cjki.org/cjk/reference/Lexicographic_criteria_Halpern.pdf> on 13 January 2022.

Halpern, J. (2020). Enhancing Arabic Speech Technology with Comprehensive Arabic Training Lexicon. Downloaded from <https://www.cjk.org/wp-content/uploads/TTS_Report.pdf > on 07 August 2023.

Haugen, E. (1972). Schizoglossia and the Linguistic Norm, in *Studies by Einar Haugen*, Berlin, Boston: De Gruyter Mouton, pages 441–445. doi: https://doi.org/10.1515/9783110879124.441

International Phonetic Association (1999). Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press. doi: https://doi.org/10.1017/9780511807954

Koehn, P. (2020). Neural Machine Translation. Cambridge University Press, Cambridge, UK. doi: https://doi.org/10.1017/S1351324920000650

Koperniak, S. (2017). Artificial data give the same results as real data — without compromising privacy. Institute for Data, Systems, and Society. Retrieved from <https://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303>

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt. https://www.researchgate.net/publication/228693973_The_penn_arabic_treebank_Building_a_large-scale_annotated_arabic_corpus

Mejdell, G. (2014). Luġat al-ʾumm and al-luġa al-ʾumm - the 'mother tongue' in the Arabic context, in *Arabic and Semitic Linguistics Contextualized*, Harrassovitz, pages 214–226. https://doi.org/10.2307/j.ctvc2rmgq.16

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth. R.M. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf

Ryding, K.C. (2005). A Reference Grammar of Modern Standard Arabic, Cambridge University Press, Cambridge, UK. doi: https://doi.org/10.1017/CBO9780511486975

Smrž, O. (2007). ElixirFM — Implementation of Functional Arabic Morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to*

Semitic Languages: Common Issues and Resources, pages 1–8, Prague, Czech Republic, June. Association for Computational Linguistics. https://aclanthology.org/W07-0801

Soudi, A., Eisele, A. (2004). Generating an Arabic Full-form Lexicon for Bidirectional Morphology Lookup. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/567.pdf

Taji, D., Khalifa, S., Obeid, O., Eryani, F., and Habash, N. (2018). An Arabic Morphological Analyzer and Generator with Copious Features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 140-150, Brussels, Belgium, October. Association for Computational Linguistics. https://aclanthology.org/W18-5816/

Wells, J.C, (1997). SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B. https://www.phon.ucl.ac.uk/home/sampa/

Tahon, M,, Qader, R,, Lecorvé, G,, Lolive, D. 2016. Improving TTS with corpus-specific pronunciation adaptation. *Interspeech*, San Francisco, United States https://www.isca-archive.org/interspeech_2016/tahon16_interspeech.pdf