Developing Full-Form Lexicons for Arabic and Its Dialects: The Case of ArabLEX as a Framework for Dialectal Lexicon Compilation

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

Yannis Haralambous
IMT Atlantique & UMR CNRS 6285 Lab-STICC
Technopôle 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 data. This presents a challenge for Arabic due to its abundant morphological, orthographic, and phonemic variation in both MSA and its various dialects. 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 an exhaustive listing of forms with detailed descriptions that can significantly aid in mitigating the inherent ambiguity of Arabic. DiaLEX, a companion of ArabLEX, covers dialects such as Egyptian, Emirati, and Hijazi and shares much of the same attributes. Since both resources use largely the same compilation methodology, they serve as a methodological framework for creating comprehensive Arabic lexical resources and full-form lexicons for dialects.

1 Introduction

1.1 What is a Full-Form Lexicon

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 lemma. 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, the abundance of combinatorics (stem, affixes, clitics) can result in full-form lexicons with hundreds of millions of entries.

1.2 The Case of Arabic

Arabic[1] is based on roots and patterns (templatic morphology) (Ryding, 2005), so we are not just dealing with stems and affixes as in many western 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 Arabic root + pattern + clitic 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. The processing performed by morphological engines is supported by lexical databases, such as ta-

[1] Arabic refers to MSA, the official language of 380M people, but (practically) no one's mother tongue (Haugen, 1972; Mejdell, 2014) and its dialects.

bles for stems, clitics, and affixes (Sawalha, 2011). 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 word forms. A notable outlier worth mentioning is the Arabic full-form lexicon and Finite State Transducer (FST) project by Soudi and Eisele (2004), which offers a targeted solution focusing on Arabic morphology. However, ArabLEX distinguishes itself with far greater scale and versatility, supporting broader NLP and speech applications.

1.4 Introducing ArabLEX

Unlike morphological engines, ArabLEX is a standalone lexical database. It can be integrated with such engines. Its goal is to act as a comprehensive database to support morphological engines and NLP tools.

ArabLEX is a *full-form lexicon* aiming to be as comprehensive as possible. In the first phase (May 2024) ArabLEX contains about 530 million MSA 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 of these words and includes a rich set of grammatical, morphological, phonological, and orthographic attributes. This makes it suitable for NLP applications such as machine translation, named entity recognition, corpus annotation, and morphological analysis and generation. Special emphasis is placed on speech technology by providing such attributes as accurate phonemic transcriptions as well as full diacritization. It is available through ELRA.

In the following we will present ArabLEX's main features. The phonemic transcriptions in this paper are italicized and given in the CARS system (Halpern, 2009a). Transliterations are given in the Buckwalter transliteration system (Buckwalter, 2002) and enclosed in forward-slashes.

1.5 Introducing DiaLEX

In parallel to ArabLEX, a series of full-form lexicons for the major Arabic dialects called DiaLEX has been developed, based on the same methodology used for ArabLEX. A recurring problem when dealing with Arabic dialects corpora is the absence of established orthographies. The data used in the development of DiaLEX have been processed by native-speaking experts, aware of the local conventions. Nevertheless, some rules were strictly observed, such as using MSA orthography whenever the MSA graphemic sequence can be phonetically realized in the phonetic space of the dialect. In many cases, alternative spellings are given. The spelling of dialects is a rapidly evolving field subordinate to cultural, political, and religious factors. DiaLEX's approach is fundamentally descriptive and prescriptive. 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).

The compilation methods employed to create ArabLEX and DiaLEX can serve as a methodological framework for creating Arabic full-form lexica, thus paving the way for creating Arabic language resources that are both accurate and comprehensive.

While this paper mainly focuses on ArabLEX, the linguistic challenges and compilation methods largely apply to Arabic dialects and DiaLEX as well. The strategies utilized to compile ArabLEX and DiaLEX serve as a methodological framework for creating comprehensive Arabic lexical resources and full-form lexicons for other dialects, such as Palestinian.

This paper focuses on ArabLEX. While the linguistic challenges in dialectical Arabic often differ, the compilation methods on the whole apply to DiaLEX as well.

2 Levels of Ambiguity

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 *ātíhima̱* to the stem *kā́tib* 'writer' yields the complex form *walika̱tibātíhima̱* (وَلِكَاتِبَاتِهِمَا) 'and to their (two) female writers'. This kind of combinatorics results in 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 كتبت is morphologically ambiguous (‘I wrote,’ ‘you (feminine) wrote,’ etc., all forms belonging to lexeme ‘to write’) while كتب is also lexically ambiguous (‘he wrote’: lexeme ‘to write,’ ‘books’: lexeme ‘book’).

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

2.2 Orthographic Disambiguation

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 (Boumaraf et al., 2022), 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) *ʔalif alfaaSila* (otiose *alif*) (Ryding, 2005), orthographic conventions not being pronounced (e.g., كتبوا being realized as *katabu*), (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.

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

2.3 Word Stress and Vowel Neutralization

Ambiguity can affect not only linguistic analysis but also speech synthesis, and in particular, prosody (word stress) and vowel neutralization, which play a critical role in ensuring that synthesized speech sounds natural. To take an example, نا *naa* is written as a long vowel in أنا but is shortened to *na* when uttered. This complex issue is described in detail in Halpern (2009c).

3 Speech Technology

3.1 Arabic Speech Technology

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 to be mispronounced, especially cliticized words. In this survey, 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*	**ɛádada*	**ɛádada*	*ɛáddada*
الكاتب	ٱلْكَاتِبُ	**lkā́tibi*	*lkā́tibu*	*lkā́tibu*	*lkā́tibu*
ما	مَا	*ma̱*	*ma̱*	*ma̱*	*ma̱*
الحكام	ٱلْحُكَّامَ	**lḥukkā́mi*	**lḥukkā́mi*	**lḥukkā́mi*	*lḥukkā́ma*

Table 1: Mispronunciations in composed text

ArabLEX addresses these shortcomings by serving as a comprehensive pronunciation dictionary to enhance the quality of both text to speech (TTS) and automatic speech recognition (ASR). It includes an NLP-oriented morpho-phonemic transcription called CARS (Halpern, 2009a), which accurately represents Arabic phonemes, *while also encoding morphological information* such as vowel neutralization. In addition, two phonetic transcriptions – SAMPA (Wells, 1997) and IPA (International Phonetic Association, 1999) – can be used to ensure accurate phonetic realizations.

3.2 ASR Accuracy

ASR systems must recognize alternative pronunciations, including informal ones. For example, the standard pronunciations of كاتبون ‘writers’ and أكتب ‘I write’ are *ka̱tibū́na* and *ʾáktubu,* but the less formal variants *ka̱tibū́n* and *ʾáktub* are very widespread.

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 and dialects. For example, رَجَعْتُ إِلَى ٱلْبَيتِ 'I returned home', pronounced *rajáεtu ʾĭla̠‿lbáyti,* in pausal form becomes *rajáεtu ʾĭla̠‿lbayt* and in spoken MSA becomes *rajáεt ʾĭla̠‿lbayt.* Note how the final *ti* and *tu* are truncated to *t.*

The example above is 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 includes the IPA representation for the standard MSA, namely [dʒ] for /j/, but will also include the regional allophones [ʒ] and [g] in a future edition.

The availability of phonetic transcriptions is particularly relevant, as phonetics can be utilized to improve ASR systems (Feng et al. 2023).

4 Machine Translation

Although Neural Machine Translation (NMT) has dramatically improved translation quality, it has some shortcomings (Koehn, 2020). Some issues in Arabic are (1) the high orthographic ambiguity, (2) the morphological complexity (forms like ولكتاباتهما 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. Additionally, the proper noun modules of ArabLEX, representing the most comprehensive collection of native and foreign proper nouns to our knowledge, areknowledge, 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).

5.2 ArabLEX Compared to Other Resources

Previous efforts to compile extensive Arabic lexicographical or lexical databases have yielded datasets containing around 200,000 unique lemmata. These datasets tend to lack a diverse set of attributes. By contrast, detailed datasets typically contain around 30,000 unique lemma entries, e.g., the CALIMA dataset for Egyptian Arabic (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 of its counterparts. Especially by offering phonetic (IPA, XSAMPA) and phonemic (CARS) transcriptions and fully diacriticized Arabic, ArabLEX fills a gap in current lexical resources.

Another advantage of ArabLEX is the total number of entries accessible for explicit analysis; that is, entries that are pre-generated as opposed to on-the-fly. For example, the CALIMA dataset contains approximately 48 million entries that can be obtained when all supported word forms are exhaustively generated (AlShuhayeb, 2023). By contrast, ArabLEX consists of 530 million pre-compiled entries, 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. To illustrate this, we compared some features of ArabLEX and CALIMA Star ("Calima" below), the most advanced morphological engine (as of May 2024), 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.

For example, the cliticized forms كَتَبْتُنِي, كَتَبْتُنَا 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 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) ArabLEX takes great care to include only grammatically valid forms. Calima, on the other hand, generates 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 اُكْتُبْ, اُكْتُبِي, etc. are not generated even when explicitly requested via the user interface.

5.4 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, root, 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

5.5 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**
Diacriticized	مُحَمَّدٌ
Phonemic	*muhammadun*
Phonetic	[muˈħɛ̈mmɛ̈dun]
X-SAMPA	mu"X\E_"mmE_"dun
Transliterated	muham~adN

Table 4: Phonological attributes for محمد

5.6 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	*walikātibíkuma̱*
Lemma	كاتب	*kā́tibun*
Stem	كاتب	*kā́tib*
Proclitic	ول	*wali*
Enclitic	كما	*(i)kúma̱*
Root	ك-ت-ب	*k-t-b*

Table 5: Moprhological attributes

Orthographic attributes are useful for orthographic disambiguation, which is necessary for word and entity recognition, TTS, morphological analysis, 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.

5.7 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, 2009b).

6 Compilation Methods

6.1 Quality Control

It can be argued that generating entries by rules and templates can result in a large number of nonexisting or erroneous forms. Extreme care has been taken to ensure that only grammatically and, as far as possible, semantically valid forms are included.

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 a data validation process used by the ArabLEX team to refine the vocalization validation module (VBW_INTEG) and ensure accurate, fully vocalized Arabic and phonemic transcriptions for speech technology:

(1) A program validates correct vocalization of inflections, 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 proofreaders, who manually classify, analyze, and rectify them. (4) Based on the feedback of proofreaders, the validation rules are then either adjusted or the 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.

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/.

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, hand-vetted conjugations 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 enclitisize, 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/, /wa-bayotN/, 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

7 Future Work

The development of ArabLEX and DiaLEX is continuous. Planned expansions include technical terms, named entities, phonological attributes,

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

orthographic variants, alternative pronunciations, and additional word classes (POS). Especially noteworthy are new headwords that consist of multiword expressionsExpansions of ArabLEX will continue, 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 إِيلَام 'more painful'), negative elatives (with أَقَلُّ or أَخَفُّ), inflected numerical expressions, phrasal verbs, compound tenses, verb negation, and more.

The addition of clitics and inflections lead to ArabLEX exceeding 500 million records (15 billion data points). It is expected to reach about one billion records in the future.

Likewise, DiaLEX will be expanded to include both broad and narrow phonological transcriptions, along with more orthographic and phonetic variants. Additionally, the Palestinian dialect will be added (PA_LEX). A paper detailing the specific challenges and methodological differences between ArabLEX and DiaLEX is forthcoming.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.

References

A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak. 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, San Diego, CA*, pages 11–16. Association for Computational Linguistics.

F. Alshargi, S. Dibas, S. Alkhereyf, R. Faraj, B. Abdulkareem, S. Yagi, O. Kacha, N. Habash, and O. Rambow. 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*, pages 137–147.

H. AlShuhayeb, B. Minaei-Bidgoli, M. E. Shenassa, and S. Hossayni. 2023. Noor-Ghateh: A Benchmark Dataset for Evaluating Arabic Word Segmenters in Hadith Domain. ArXiv preprint.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

M. Boudchiche, A. Mazroui, M. Ould Abdallahi Ould Bebah, A. Lakhouaja, and A. Boudlal. 2017. AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. *Journal of King Saud University – Computer and Information Sciences*, 29(2):141–146.

A. Boumaraf, S. Bekal, and J. Macoir. 2022. *The Orthographic Ambiguity of the Arabic Graphic System: Evidence from a Case of Central Agraphia Affecting the Two Routes of Spelling*. Behavioural Neurology.

T. Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania. *2002*, LDC Catalog No.: LDC2002L49.

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

CJK and The Dictionary Institute. 2011. The CJKI Arabic Verb Conjugator.

N. Gadelli. 2015. *The morphological integration of loanwords into Modern Standard Arabic: Towards a morphological categorization of loanwords*. Ph.D. thesis, Lund University, Sweden.

N. Habash, O. Rambow, and R. Roth. 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.

J. Halpern. 2009a. CJKI Arabic Romanization System (CARS).

J. Halpern. 2009b. 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.

J. Halpern. 2009c. 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.

J. Halpern. 2019. Lexicographic Criteria for Selecting Multiword Units for MT Lexicons.

J. Halpern. 2020. Enhancing Arabic Speech Technology with Comprehensive Arabic Training Lexicon.

E. Haugen. 1972. Schizoglossia and the Linguistic Norm. In *Studies by Einar Haugen*, pages 441–445. De Gruyter Mouton, Berlin, Boston.

P. Koehn. 2020. *Neural Machine Translation*. Cambridge University Press, Cambridge, UK.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt*, pages 102–109.

G. Mejdell. 2014. Luġat al-ʾumm and al-luġa al-ʾumm - the 'mother tongue' in the Arabic context. In *Arabic and Semitic Linguistics Contextualized*, pages 214–226. Harrassovitz.

A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth. 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.

K. C. Ryding. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press, Cambridge, UK.

M. S. S. Sawalha. 2011. *Open-source resources and standards for Arabic word structure analysis: Fine-grained morphological analysis of Arabic text corpora*. Ph.D. thesis, The University of Leeds School of Computing.

O. Smrž. 2007. ElixirFM — Implementation of Functional Arabic Morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic*, pages 1–8. Association for Computational Linguistics.

A. Soudi and A. Eisele. 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*. European Language Resources Association (ELRA.

M. Tahon, R. Qader, G. Lecorvé, and D. Lolive. 2016. *Improving TTS with corpus-specific pronunciation adaptation*. Interspeech, San Francisco, CA.

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

J. C. Wells. 1997. SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter. Part IV, section B, Berlin and New York.