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

Abstract

Various factors contribute to the difficulties of Arabic natural language processing (NLP), posing special challenges in the areas of speech technology, machine translation (MT), and morphological analysis. NLP methods based on statistical and algorithmic methods, and even methods using neural networks and deep learning, are inadequate by themselves, as evidenced by the relatively low quality of applications such as speech technology and named entity recognition (NER). To enhance these methods, the most advanced Arabic NLP systems have adopted lexicon-driven approaches to supplement systems based on neural networks and deep learning. This paper describes ArabLEX, a *full form lexicon* specifically designed for Arabic NLP and speech technology. While normal dictionaries are limited to canonical forms, full form lexicons aim to explicitly list all inflected, conjugated, declined, and cliticized forms that occur in a language. Already deployed by major IT companies, ArabLEX is the most comprehensive Arabic computational lexicon ever created. It currently covers over 530 million general vocabulary and proper noun entries with a rich set of grammatical, morphological, phonological and orthographic attributes.

Keywords: Arabic NLP, full form lexicon, speech technology, Arabic morphology, large resources

1. Introduction

1.1 Why a Full Form Lexicon

Wordform, as opposed to *lexeme*, refers to an inflected, conjugated, declined or cliticized form that is a member of a lexeme class. In English, for example, the lexemes *eat* and *boy* (canonical forms) have the following members, namely *eat, eats, eating, eaten, ate* and *boy, boys, boy's, boys'* respectively. While dictionary headwords are normally limited to canonical forms, a *full form lexicon* aims to explicitly include all members of a lexeme class.

Traditionally, MT and other NLP applications have been based on rules or statistical models. In recent years, neural machine translation (NMT) is becoming the norm. Despite its remarkable achievements, when it comes to Arabic this technology has some shortcomings, such as the handling of proper nouns (Halpern, 2009) and multiword expressions (MWE) (Halpern, 2019), inadequate bilingual training corpora, and relatively low accuracy rates.

A full form lexicon can enhance NLP technology in various ways, especially by providing detailed morphological and phonological information for all wordforms explicitly, rather than relying on real-time algorithmic generation and analysis.

Note that all phonemic transcriptions in this paper are given in the CARS system (Halpern, 2009), designed by our institute for pedagogical and speech applications.

1.2 Introducing ArabLEX

ArabLEX stands for "Comprehensive Arabic Full Form Lexicon." Currently (January 2022) it contains about 530 million entries 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 explicit and exhaustive coverage of all inflected, declined, conjugated and cliticized forms, and includes a rich set of grammatical, morphological, phonological and orthographic attributes for each wordform. ArabLEX is designed to support NLP applications such as morphological analysis, machine translation, named entity recognition (NER), and morphological generation. Special emphasis is placed on speech technology by providing accurate phonemic and phonetic transcriptions as well as full vocalization for each entry.

The database is undergoing constant maintenance and expansion, and during 2022 it is expected to exceed one billion entries. To our knowledge, it is the most comprehensive Arabic computational lexicon ever created.

1.3 Previous Work

Many Arabic morphological modeling tools have been developed over the last few decades. These tools are designed for such tasks as morphological analysis, tokenization and generation of inflected and conjugated forms, POS tagging, and disambiguation. Below we will refer to such tasks as analysis and generation, and to such tools as morphological engines. The main well known tools include MADA (Habash, Rambow, and Roth, 2009), BAMA (Buckwalter, 2002), PATB (Penn Arabic Treebank) (Maamouri 1., 2004), FARASA (Abdelali et al., 2016), MADAMIRA (Pasha et al., 2014), and Elixir_FM (Smrž, 2007) . A more recent, and no doubt the most ambitious of these tools, is CALIMA-Star (Taji et al., 2018).

Despite the high performance of some of these tools, as Taji et al. (2018) pointed out, they do have various shortcomings, such as not being fully consistent, ignoring lexical rationality, lack of or inadequate phonological attributes, and more. A common feature of these tools is that they are *morphological engines,* designed to perform analysis and generation, as defined above. Naturally, the processing performed by these engines is supported by lexical databases, such as tables for stems, clitics and affixes (Halpern, 2018), but the goal of these tools is to perform computational tasks, not to function as lexicons in their own right.

In contrast, ArabLEX is not a morphological engine; it is a stand-alone module that can be integrated into the morphological engine but it does not do any

computational task on its own. Its goal is to act as a comprehensive, highly structured database to support morphological engines and the development of NLP tools. In theory, the engine can query the lexicon as an external module by function call or API, but in practice it is desirable to integrate it into the engine itself. To use an analogy, if we liken a morphological engine to the engine of a car, then a full-form lexicon is like the fuel that drives the engine, but it is not the power that drives the car itself (e.g. a TTS application).

ArabLEX is thus different from morphological engines in that it acts as a *static* comprehensive source of full form lexical data to power the engine. The morphological engine, on the other hand, is the dynamic module that performs the actual analysis and generation tasks. The engine and lexicon are not in competition; they have distinct roles and are meant to work in synergy.

2. Morphological Ambiguity

Arabic has a rich and ambiguous morphology. Inflection is indicated by changing the vowel patterns as well as by affixation and cliticization (templatic morphology). Not only can words be inflected, declined and conjugated ("inflected" for short), they can also take many proclitics and enclitics. For example, adding the proclitics *wa* 'and', *li* 'to', and the enclitic *ātíhimą* to the stem *kắtib* 'writer' yields the complex form *walikątibātíhimą* (وَلِكَاتِبَاتِهِمَا) 'and to their (dual) female writers', This results in a very large number of wordforms. For example, the full paradigms for كَاتِبٌ *kātibun* 'writer' and كَتَبَ *kataba* 'write' reach about 5,660 and 6,900 forms respectively (by comparison Japanese verbs have about 2,000 forms).

The morphological complexity of such forms as *walikątibātíhimą*, and the absence of vowel diacritics, makes Arabic morphological analysis and speech technology especially challenging. That is, determining the morphological composition of such forms and disambiguating the vowels require techniques that are often beyond the capabilities of state-of-the-art Arabic NLP technology.

3. Orthographic Ambiguity

3.1 Why is Arabic Ambiguous

One reason that Arabic NLP, especially speech technology, lags behind other major world languages is that the Arabic writing system is highly ambiguous. Many factors contribute to a high level of orthographic ambiguity, as described in Halpern (2009).

Conventional wisdom has it that Arabic is ambiguous "due to the non-representation of short vowels." In fact, a whole gamut of factors contribute to ambiguity, posing major challenges to Arabic speech technology (Halpern, 2002). This includes, among others, (1) the absence of short vowels (e.g. كاتب represents the seven wordforms *kātib, kātibun, kātibin, kātaba, kātibi, kātiba, kātibu,* (2) long *ā* is represented by ا as in سوريا or by آ as in آسيا, but some bare alifs represent *tanwiin* rather than long *ā*, as in شـكرا *shukran*, (3) *'alif alfaaSila* (otiose alif) (Ryding, 2005), an orthographic convention not pronounced, e.g., كتبوا is realized as *katabu*[1]), (4) the omission of s*hadda* indicating consonant gemination, e.g., محمـد (vocalized مُحَمَّد) provides no clues that the /m/ is doubled, and (5) vowel neutralization is sometimes lexically determined and thus cannot be predicted from the orthography; e.g., فـي القـاهرة 'in Cairo' is pronounced *fi-lqaahira*, not *fii-lqaahira*.

3.2 Orthographic Disambiguation

A central issue in Arabic NLP, especially speech technology, is identifying which of the possible wordforms an ambiguous string like كاتباتك (vocalized as كَاتِبَاتِـكِ) represents. This can represent any of six wordforms, each with a different meaning, a different pronunciation, or a different morphological or syntactic function.

The process of identifying the correct form in context is referred to as orthographic disambiguation (Halpern, 2008). The rich set of grammatical and morphological attributes in ArabLEX can help train the language model to correctly disambiguate such forms. That is, these attributes provide the grammatical and morphological context in which ambiguous strings occur, which helps determine the correct wordform for that context and thus the correct pronunciation as well. The attributes for كَاتِبَاتِكِ show that it refers to plural female writers in the definite state, who belong to second person singular feminine in the genitive case, which helps to determine the correct pronunciation of *kątibắtiki.*

4. Enhancing Speech Technology

4.1 Arabic Speech Technology

Though recent advances in neural networks have dramatically improved the quality of speech technology, a recent survey comparing the TTS provided by major IT players showed that Arabic significantly lags behind other major languages (Halpern, 2020).

ArabLEX addresses these shortcomings by serving as a comprehensive pronunciation dictionary for enhancing the quality of both TTS (text-to-speech) and ASR (automatic speech recognition). To that end, it includes an NLP-oriented phonemic transcription called CARS (Halpern, 2009) and two phonetic transcriptions (SAMPA and IPA) whose goal is to support the training of TTS and ASR models.

For example, the entry وَلِكَاتِبَاتِهِمَا has a CARS transcription of *walikątibātíhimą*, an accurate phonemic representation. It consists of the stem *kắtib* 'writer' combined with the proclitics *wa* 'and' and *li* 'to' and the enclitic *ắtíhimą* 'their'. The *ą* indicates two occurrences of the long vowel *ā* that have been neutralized to short *ą*, indicated by the underline.

4.2 Improving TTS Accuracy

The extreme orthographic ambiguity of Arabic has led to unacceptably high error rates in TTS, even by the major IT players. The CJKI survey (Halpern, 2020) revealed that

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

it is not unusual for over 50%, and even 80%, of the words in a sentence, especially for cliticized words, to be mispronounced. For example, the cliticized word وَلِلْكَاتِبِينَ, correctly pronounced *walilkatibīna*, is mispronounced as *walilkātibáyna*. In Table 1, pronunciation errors are shown by the asterisk.

Unvoc-alized	Vocal-ized	Google (13%)	iOS (31%)	Bing (25%)	CJKI
عدد	عَدَّدَ	*ɛádadu**	*ɛádada**	*ɛádada**	*ɛáddada*
الكاتِ	اَلْكَاتِبُ	*lkắtibu*	*lkắtibi**	*lkắtibu*	*lkắtibu*
ما	مَا	*mạ*	*mạ*	*mạ*	*mạ*
الحكام	اَلْحُكَّامَ	*lḥukkắmi**	*lḥukkắmi**	*lḥukkắmi**	*lḥukkắma*

Table 1: Mispronounced words in composed text

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 pronounced by Google TTS as follows:

ɛádadu [ɛáddada] lkắtibu mạ qắla ʾínna hạ́ʾulắʾi lḥukkắmi* [lḥukkắma] yafɛalūnahu fī lkhắriji míthli* [míthla] lhajamắti lʾilikturụ́niyyati wamuṭārádati lmuɛariḍ́na llajiʾ́na fī lɛawắṣimi lgharbíyyati.*

Words followed by an asterisk are pronounced incorrectly with the correct pronunciation following in brackets.

A more recent test (December 2021) on تَصْحَبُوا using the Google engine produced *taṣḥábụwa*, instead of the correct *táṣḥabụ*. Note that 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 shows that the TTS error rate for the major engines is unacceptably high. The error rates in composed texts ranged from 13% to 25%, whereas in web-extracted texts they ranged from 70% to 90%.

4.3 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 in actual pronunciation to *na*. This is a complex issue, described in detail in Halpern's paper on Arabic stress (2009).

The phonemic and phonetic transcriptions in ArabLEX explicitly indicate precise word stress and vowel neutralization for each entry. For example, in the IPA [wɛlikɛ̈ːˈtibikumɛ(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.

4.4 Improving ASR Accuracy

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

Such alternatives include pausal forms and final vowel elision. The former refers to sentence final forms causing final vowels (case endings and nunation) to be elided in Classical Arabic, while the latter is the elision of certain final vowels in both medial and final forms, very common in spoken MSA. For example, رَجَعْتُ إِلَى ٱلْبَيتِ 'I returned home', whose standard pronunciation is *rajáɛtu ʾílạ‿lbáyti*, in pausal form becomes *rajáɛtu ʾílạ‿lbayt* and in spoken MSA becomes *rajáɛt ʾílạ‿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 such words as *jamal* 'camel' is pronounced [g] in Egypt, [dʒ] in the Gulf region, and [ʒ] in the Levant. These are not *dialectal* pronunciations, but *regional* variants of MSA as spoken in those regions. 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 exhaustively includes transcriptions of the most common non-standard forms and regional allophones.

5. Enhancing Machine Translation

Although neural machine translation has achieved dramatic improvements in translation quality, it does have some shortcomings, as pointed out by Philipp Koehn (2020) and Halpern (2018). 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 (which are often cliticized), and (4) the large number of wordforms for Arabic nouns and verbs.

ArabLEX can significantly enhance the accuracy of Arabic MT. Since it provides comprehensive coverage of inflected and cliticized forms, it can be used as a pseudo-corpus to train the language model and enable more accurate morphological, syntactic, and semantic analysis. In addition, the proper noun modules of ArabLEX – DAN, DAF and DAP – are bilingual and romanized so they can serve as a bilingual dictionary.

When NMT first appeared, it was believed that lexicons could not be integrated into NMT systems (Halpern, 2018). Later it was shown that they can be by regarding a lexicon as a kind of sentence-aligned, parallel corpus and assigning a higher probability to lexicon lookup results to override the normal NMT algorithms. By using such techniques, it is possible to integrate ArabLEX into Arabic NMT systems (Halpern, 2018).

6. ArabLEX in Action

6.1 Scope and Coverage

The first release of ArabLEX in 2021 covered about 530 million full form entries (cross-referenced to their lemmas) in the domains of general vocabulary and proper nouns.

ArabLEX consists of the following four main modules:

DAG	Arabic General Vocabulary	83 million
DAN	Arabic Names	218 million
DAF	Arabic Foreign Names	226 million
DAP	Arabic Place Names	6 million

Table 2: ArabLEX coverage

ArabLEX has 30 data fields with detailed grammatical, phonological, morphological and orthographic attributes, described below and in more detail in Halpern (2020).

In the initial version, general vocabulary is limited to the core vocabulary of Arabic and some medium-frequency headwords. The reason for the hundreds of millions of entries is the exhaustive coverage of clitics, especially proclitics, and comprehensive inclusion of cliticized personal names. In future versions, more general vocabulary and high frequency technical terms will be added.

It can be argued that generating entries by rules and templates can result in a large number of non-exiting 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 exist, or have not been observed to exist, they are indeed valid and could occur in the future. For most applications, nonexistent forms have no negative effects, whereas missing entries do.

6.2 Grammatical Attributes

The grammatical attributes are useful for morphological analysis, orthographic disambiguation, POS tagging, semantic analysis, and more. These include such attributes as codes for gender, number, case endings and person, as well as the stem, definiteness, lexical rationality and the lemma. The main grammatical attributes are shown in Table 3.

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

6.3 Phonological Attributes

The phonemic and phonetic transcriptions are useful for training speech technology models, both TTS and ASR. Phonological attributes include precise fully vocalized Arabic with accurate phonemic and phonetic transcriptions as well as word stress and vowel neutralization for all entries. The main phonological attributes are shown in Table 4.

Data Field	Value
Vocalized	مُحَمَّدٌ
Phonemic	*muhammadun*
Phonetic	[muˈħɛ̈mmɛ̈dun]
X-SAMPA	mu"X\E_"mmE_"dun
Transliterated	muham~dN

Table 4: Phonological attributes for محمد

6.4 Morphological Attributes

The morphological attributes include all inflected, conjugated, declined, and cliticized wordforms, 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	وَلِكَاتِبِكُمَا	*walikātibikumạ*
Lemma	كَاتِبٌ	*kắtibun*
Stem	كَاتِب	*kắtib*
Proclitic	وَلِ	*wali*
Enclitic	كُمَا	*(i)kúmạ*
Root	ك-ت-ب	*k-t-b*

Table 5: Morphological attributes

6.5 Orthographic 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 attributes include orthographic variants for both vocalized and unvocalized Arabic, including pausal and elided forms, and even common typographical errors. Typical orthographic variants are shown in Table 6.

Data Field	Value
Variant 1	ألكسندرة
Variant 2	الكسندرة
Variant 3	ألكسندره

Variant 4	الكسندره
Variant 5	ألكسندرا
Variant 6	الكسندرا

Table 6: Orthographic variants for *Alexandra*

As can seen above, ه and ة are sometimes interchangeable in names. Orthographic variants also include what can be considered to be *allographs,* as 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.

6.6 Named Entity Recognition

The DAN module of ArabLEX is derived from the Database of Arabic Names (DAN) (Halpern, 2009), which covers about 100,000 vocalized personal names and their 6.5 million romanized variants (romanized variants are not included in DAN). 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.

6.7 Accessing ArabLEX

As mentioned in section 1.3, since ArabLEX is not a morphological engine, it does not actually *do* anything unless coupled to an engine. However, we are developing a lookup tool, called ArabFIND*,* that enables access through a user-friendly interface, a command line query, or an API. This tool also does some intelligent fuzzy matching and normalization when strings are not found by exact match. Since this is only a lookup tool, it does not perform analysis, disambiguation, or generation, but returns candidate(s) that match the query parameters, which effectively is equivalent to analysis and generation. If the parameter consists of an inflected form, it will return a list of matching attributes, equivalent to analysis. If the parameters consist of a lemma combined with various attributes, it will generate all forms specified by those parameters.

Since ArabFIND basically only performs (intelligent) lookup, the algorithms are significantly simpler than in morphological engines. In theory, the morphological engine can query ArabLEX in real time using ArabFIND, but for most engines it is probably more efficient to integrate the lexicon into the engine itself.

Parameter	Value
Form	وَلكاتبَاتهمَا
Mode	analysis
Output fields	pos, gen, num, case, per, proc, lemma, enc

Table 7: Input parameters for analysis

No	pos	gen	num	case	per	proc	lemma	enc
1	N	F	P	GEN	3DM	وَل	كاتبٌ	اتهمَا
2	N	F	P	GEN	3DF	وَل	كاتبٌ	اتهمَا
3	N	F	P	ACU	3DM	وَل	كاتبٌ	اتهمَا
4	N	F	P	ACU	3DF	وَل	كاتبٌ	اتهمَا

Table 8: Partial analysis of وَلِكَاتِبَاتِهِمَا *walikatibātihimā*

As can be seen in Table 7, the cliticized form وَلِكَاتِبَاتِهِمَا *walikatibātihimā* is passed to the tool along with several parameters, and the 'output fields' parameters indicate which fields to output. The results consist of four candidates of the attributes requested by the parameters. The analysis indicates that the requested form is a noun, the feminine plural of the lemma كَاتِبٌ in the genitive or accusative case belonging to two people in the third person dual masculine or feminine.

7. Compilation Methods

The techniques used to compile ArabLEX are manifold and complex, and will be the subject of another paper. Briefly, for about a decade, our team of lexicographers and specialists in morphology, phonology and computational lexicography engaged in the development of Spanish (Carbonell et al., 2006), Japanese and Arabic full form lexicons, which have significantly contributed to several major NLP projects. This was followed by expansion, proofreading and validation of ArabLEX, with the aim of covering all inflected and cliticized wordforms for both general vocabulary (Halpern, 2020) and proper nouns (Halpern, 2009).

The raw data was collected from various lexicographic sources and corpora (Halpern, 2016), which underwent extensive semi-automated validation (sanity checks), semi-automated diacriticization, and human proofreading. Verb generation was based on our CAVE system (CJKI Arabic Verb Conjugator) (Halpern, 2011) using thoroughly validated verb paradigm templates. Cliticization, declension and inflection were based on about 100 clitic templates that define constraints on which enclitics can combine with which proclitics for different categories of stem types, inflectional endings, and words classes (POS). For example, the subset of clitics for personal names differs from that of ordinary nouns. Table 9 presents a snippet of a typical template using Buckwalter transliteration (2002):

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

Table 9: Template for nouns that end in /pu/

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 10: Possible combinations of clitics

In the case of شَرْعِيَّة *sharεíyyat*, the template indicates that the enclitic *uki* and the proclitic *fa* can be attached to yield فَشَرْعِيَّتُكِ *fasharεíyyatuki*. Algorithms are then used to generate the fully cliticized forms for each wordform based on the appropriate template. Thus the clitics are not merely blindly concatenated to the base form -- there are thousands of orthographic (liaison), grammatical and semantic constraints on clitic combinatorics. The templates and their accompanying algorithms precisely define which clitics can be combined with which base forms and the orthographic changes that occur in the concatenation process.

8. Future Work

ArabLEX is a comprehensive computational lexicon with a rich set of features designed to enhance Arabic NLP applications such as MT, orthographical disambiguation, morphological analysis, and entity recognition. For speech technology, its provides precise phonemic and phonetic transcriptions, including word stress and allophonic variants.

We will continue to expand ArabLEX by adding new entries and new features, such as more phonological attributes, technical terms and named entities. 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 إِيلَامًا 'more painful'), negative elatives (with أَقَلُّ or أَخَفُّ), fully inflected numerical expressions, phrasal verbs, compound tenses, verb negation, and more. We will also add new features to ArabFIND, a user-friendly interface with an API for quick access. In parallel, we are developing a series of full form lexicons for the major Arabic dialects, called *DiaLEX*, based on the same methods used for ArabLEX.

Since ArabLEX fully covers proclitics, enclitics and inflections, it has grown to over 500 million records (15 billion data points). As a result of the new expansions we expect it to reach about one billion records during 2022. As we have seen, morphological engines and full form lexicons are not in competition; they have distinct goals and are meant to work in synergy. ArabLEX aims to serve as the ultimate resource for the development of Arabic NLP applications.

9. 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.

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).

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.

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.

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.

Habash, N.Y. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers, San Rafael, CA, USA.

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).

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.

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.

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

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).

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.

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

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.

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).

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

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.

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).

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.

10. Language Resource References

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