Behavior Research Methods (2025) 57:104
https://doi.org/10.3758/s13428-024-02560-8

ORIGINAL MANUSCRIPT

SUBTLEX-AR: Arabic word distributional characteristics based on movie subtitles

Sami Boudelaa[1] · Manuel Carreiras[2,3,4] · Nazrin Jariya[1] · Manuel Perea[5,6]

Accepted: 18 November 2024 / Published online: 26 February 2025

Abstract

This article presents SUBTLEX-AR, a digital database providing an extensive collection of attributes related to Modern Standard Arabic words (Arabic for short). SUBTLEX-AR combines a novel dataset of 120 million word tokens from movie subtitles with 40 million tokens from newspaper articles originally collected in ARALEX (Boudelaa & Marslen-Wilson, *Behavior Research Methods, 42*, 481–487, 2010), ensuring comprehensive coverage. SUBTLEX-AR provides information about the statistical properties of Arabic words at the orthographic, phonological, morphological, and semantic levels. The database also includes information on sub-word structure properties like bigram and trigram frequencies, as well as lemmas and part-of-speech information along with their corresponding frequencies. The online interface of SUBTLEX-AR allows users either to upload a set of words to receive their properties or to receive a set of words matching constraints on predefined properties. The properties themselves are easily extensible and will be expanded over time. SUBTLEX-AR is freely accessible here: https://subtlexar.uaeu.ac.ae/

Keywords Word frequency · Morpheme frequency · Semantic similarity · Subtitles · Arabic

Introduction

Information about the distributional properties of words is an indispensable element of the researcher's toolkit in areas such as neuroscience, artificial intelligence, psychology, linguistics, and education, to name but a few. For example, surface word form frequency, that is, the number of times a particular word form occurs in a corpus, is one of the most important variables in psycholinguistics, affecting both the speed and accuracy of word recognition (Forster & Chambers, 1973; Monsell, 1991; Monsell et al., 1989; see Mandera et al., 2017, and Yonelinas, 2002, for reviews). Word frequency, together with other variables, also plays significant modulatory roles not only in processes directly related to word recognition but also in broader memory processes (e.g., short-term memory, working memory). These other variables include orthographic neighborhood (Carreiras et al., 1997; Gomez et al., 2007; Gomez & Silin, 2012; Grainger & Hannagan, 2012; Grainger & Ziegler, 2011; Grainger, 1990), phonemic frequency (Chang et al., 2023; Robson et al., 2003), syllable frequency (Carreiras et al., 1993; Carreiras & Perea, 2004; Perea & Carreiras, 1998; Gathercole & Baddeley, 1990; Stenneken et al., 2005), morpheme frequency (Ford et al., 2010; Lane et al., 2019; Taft & Zhu, 1997), and semantic distance between words (Buchanan et al., 2001; Faust & Lavidor, 2003).

Given this diverse array of word characteristics influencing language, cognitive, and memory processes, it is undoubtedly advantageous to have a unified, comprehensive, and regularly updated data source that provides researchers with information on the orthographic, phonological, and morphological frequency of words as well as their semantic properties. Here, our focus is on Modern Standard Arabic (Arabic henceforth), a language

✉ Sami Boudelaa
s.boudelaa@uaeu.ac.ae

1 Department of Cognitive Sciences, United Arab Emirates University, Al Ain 15551, UAE

2 Basque Center for Cognition, Brain, and Language, Donostia-San Sebastián, Spain

3 IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

4 University of the Basque Country, Donostia-San Sebastián, Spain

5 Universitat de València, Valencia, Spain

6 Nebrija University, Madrid, Spain

characterized by a paucity of lexical resources, with only one database that provides frequency information about words, morphemes, and graphemes. Currently, the most comprehensive database is ARALEX (Boudelaa & Marslen-Wilson, 2010), which is based on newspaper articles. Although other databases exist (e.g., Arabicorpus: https://arabicorpus.byu.edu/; the Qur'an Lexicon Project: Faizal et al., 2015; Tunisian Arabic Corpus: McNeil & Miled, 2010) and provide invaluable types of information, they are all limited in significant ways. For instance, Arabicorpus is a good source of information about surface word frequency and word collocation. However, Arabicorpus does not provide information about vowel-disambiguated items or the orthographic, phonological, morphological, or semantic structure of words. The Qur'an Lexicon Project is a unique source about word length in syllables and phonemes, as well as word frequency, lexical uniqueness point, orthographic and phonological neighborhood sizes, and orthographic and phonological Levenshtein distances among words and phonotactic probabilities. However, the Qur'an Lexicon Project is limited in scope. It is confined to 19,286 vocabulary items of classical Arabic. The Tunisian Arabic corpus operates with a small corpus that consists of 1,082,375 words and, when queried, returns little more than the query word itself in different sentence contexts. In addition, the corpus consists solely of dialect words spoken in the north of Tunisia. It thus has little to offer to researchers interested in visual word recognition processes in standard Arabic. This means that experimental research that depends on the use of linguistic stimuli in Arabic is still seriously hampered compared to other languages, particularly the Indo-European languages. This is an undesirable situation, as Arabic is the official language of 21 countries, natively spoken by 380 million people, used as the liturgical medium by more than 1.9 billion people worldwide, and is de facto the fifth most spoken language in the world. The current project is a step towards remedying this situation. More specifically, we set out to provide SUBTLEX-AR, a new database that integrates data from movie subtitles with data from newspaper articles.

In what follows, we first describe the integrative structure of SUBTLEX-AR. Second, we describe the collection and preprocessing of the movie subtitles we used to build SUBTLEX-AR. Third, we detail how we compiled the statistics pertaining to the different domains of linguistic knowledge (i.e., orthography, morphology, phonology, and semantics). Then we report a validation of SUBTLEX-AR using lexical decision data and subjective ratings by participants. We conclude by highlighting the importance of SUBTLEX-AR as an extensible, ever-improving tool likely to promote scientific progress in the field and point to how this resource will be revisited and improved over time.

The structure of SUBTLEX-AR

The terminology of the SUBTLEX-AR database derives from the recent surge in the number of lexical databases built based on movie and TV series subtitles. These are conventionally referred to as SUBTLEX, followed by a hyphen and two/three letters that designate the language in question (e.g., SUBTLEX-US: Brysbaert et al., 2012; SUBTLEX-CH: Cai & Brysbaert, 2010; SUBTLEX-NL: Keuleers et al., 2010; SUBTLEX-GR: Dimitropoulou et al., 2010; SUBTLEX-DE: Brysbaert et al., 2011; SUBTLEX-ESP: Cuetos et al., 2011; EsPal: Duchon et al., 2013; SUBTLEX-UK: van Heuven et al., 2014a, 2014b; SUBTLEX-PT: Soares et al., 2015). This new type of database completes traditional databases obtained by assembling large amounts of written texts from books and periodicals (e.g., Thorndike, 1921; Thorndike & Lorge, 1944; Kučera & Francis, 1967). We developed SUBTLEX-AR within this tradition, combining a 40-million-word corpus acquired from newspaper articles with a 117-million-word corpus from movie subtitles. The 40-million-word corpus comes from the ARALEX database (Boudelaa & Marslen-Wilson, 2010). The new component comprises 34,896 Arabic subtitle files from movies and TV shows provided by www.opensubtitles.org. These subtitles were of movies and TV shows shot over a period spanning nearly 90 years, namely from 1930 to 2020, with most movies and TV series being screened between 2003 and 2020. This corpus includes IMBD-type movies and TV series, including drama, comedy, thriller, action, crime, mystery, adventure, and romance, thus ensuring that a broad spectrum of topics and linguistic materials are sampled. The current subtitle corpus consists of 86,568 word types and 117,508,475 word tokens. This corpus size is comparable to other subtitle corpora like SUBTLEX-PT, with 132,710 word types from 78 million films and television episodes (Soares et al., 2015), and EsPal, with 244,933 word types derived from 460 million films and TV episodes (Duchon et al., 2013).

SUBTLEX-AR is freely accessible online at https://subtlexar.uaeu.ac.ae/. The first page of the database interface (Fig. 1) shows the two components in use, ARALEX and SUBTLEX, with a selectable button for each one.

As mentioned above, ARALEX corresponds to the 40-million-word corpus from newspaper articles, while SUBTLEX consists of the movie subtitle data. The user must decide which corpus to search by selecting the appropriate button. The opening page also shows a 'Choice of Display' option button that allows the user to have the query output in Arabic if checked. The query itself can be input either in Arabic or in Latin characters using the Buckwalter transliteration scheme (for examples and

Fig. 1 Screenshot of the opening page of SUBTLEX-AR

implementation of the scheme, see https://en.wikipedia.org/wiki/Buckwalter_transliteration).

Subtitle corpus preprocessing

Raw Arabic text typically comes with many superfluous features such as elongation (e.g., كبـــــير instead of كبير), partial diacritics to prime a particular reading of an orthographic string (e.g., كُتب instead of كتب), abbreviations (e.g., ف-بي-أي), symbols (e.g., @, #, etc.), and unusual orthographical forms (e.g., هيّشتا). These were removed, leaving only viable Arabic words and numbers. In a second step, we removed repeating letters (e.g., هاااااادى changed into هادى), numbers (e.g., 1, 2), and foreign characters (e.g., Persian letters چ پ ژ). Typos were identified and hand-corrected by a group of UAE University student volunteers (e.g., محفوضة corrected into محفوظة).

Beyond this, we departed from the oft-followed practice of normalizing the letters ألف *alif*, 'ا,آ,أ,إ' into 'ا', تاءمربوطة *taa marbwTp'*, 'ة, ه' into 'ة' and ياء مقصورة *yaa maqSwrp*, 'ى, ي' into 'ى' (Larkey et al., 2002). Instead, we opted to keep all these letters as they are in order to avoid losing contrasts such as forms آخذ-أخذ *I take-he took*, لكمة-لكمه *a punch-he punched him*, مجرى-مجري *stream-Hungarian*. We also retained foreign borrowings such as أكشن *action*, كاونت *count*. These preprocessing operations yielded 117,508,475 word tokens corresponding to 86,568 word types, all without diacritical marks.

To reinstate the vowel diacritics, we chose to use the Farasa diacritizer. It relies on a bidirectional long short-term memory (biLSTM) network to restore core word diacritics and case endings. It has been demonstrated to outperform other diacritizers, with an average core word diacritization error rate of 2.86% and a grammatical case ending diacritization error rate of 3.7% (Darwish et al., 2020).

SUBTLEX-AR measures

SUBTLEX-AR affords information about the orthographic, morphological, phonological, and semantic properties of Arabic words. In what follows, we describe these measures, how they were computed, and how end-users can query the database to obtain them. For each measure in each domain, the user can query the database either using a single search item or uploading a list of items, as shown in Fig. 2.

Both queries can be carried out using the Arabic script or the Buckwalter transliteration. Furthermore, for each type of measure, except semantics, SUBTLEX-AR provides two possibilities. First, the user feeds their selected items into the database and obtains the associated distributional properties. Second—and this is the more useful option for experimental researchers—the user can set up criteria for each domain to obtain the materials they need.

Fig. 2 Screenshot showing the two query options highlighted within a red rectangle

Orthographic measures

SUBTLEX-AR provides information about the following orthographic properties:

Word frequency

The term 'word' as we use it here refers to the string of letters written with white space on either side of it; as such, an Arabic word can range in structural complexity from dictionary citation forms (e.g., من *who*, كلب *dog*, خرج *go out*), through complex noun phrases (e.g., والعمل *and the work;* كأخي *like my brother*), to full sentences (e.g., أحضرته *I prepared it,* وادخلته *and I got it in*). Under the *word frequency* rubric, we provide six measures, as illustrated in Fig. 3.

Each of these measures can be individually selected by ticking the box next to it, and information about what the measure represents can be consulted by clicking on the *Help* button opposite the measure at hand. In Fig. 4, we have used the *Select All* button to query the six measures under *Word Frequency* and display their Help menus.

A particularly useful feature of SUBTLEX-AR for experimental researchers is the *Constraint to Word* option, available for all the measures under the orthography, phonology, and morphology. This feature allows the user to seek and obtain a list of words with specific properties in each domain. For example, if the user clicks on the *Constraint to Word* option in the orthography rubric, they can set the minimum and maximum values for each measure as illustrated in Fig. 5 with *Word frequency per million* and *Log Frequency*.

Figure 6 illustrates the output of this constraint-based search, which can be conveniently downloaded as a tab-delimited TXT file.

Fig. 3 Screenshot showing the statistics available for Arabic words

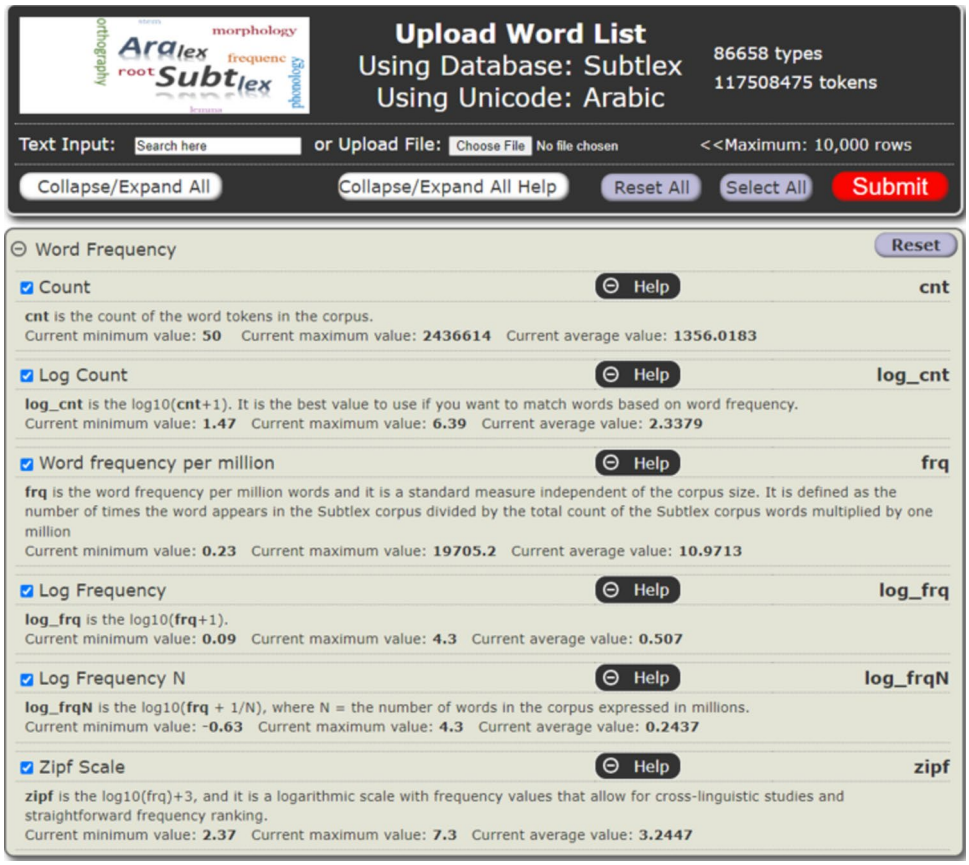
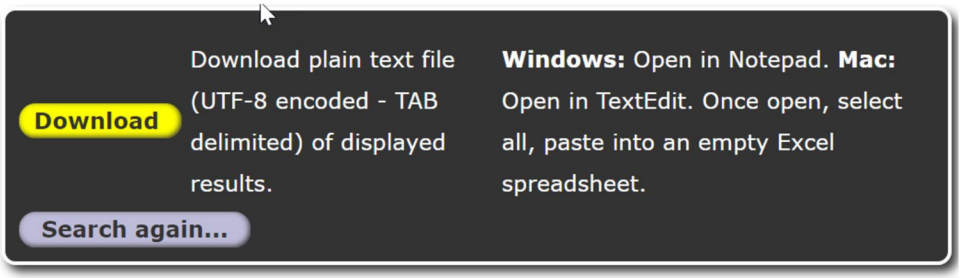


Fig. 4 Screenshot showing the measures *Count* and *Log Count* selected and their help menus displayed

Fig. 5 Screenshot showing an example of a constraint-based search for the *Word frequency per million* and *Log Frequency* measures



Token	frq	log_frq
آبي	35.14	1.56
آت	25.55	1.43
آثـار	34.59	1.56
آخذ	49.4	1.71
آخذك	20.68	1.34
آخريـن	48.31	1.7

Fig. 6 Screenshot showing a portion of the output of a constraint-based search for the *Word frequency per million* and *Log Frequency* measures under the orthography rubric

Lemma information

A lemma is the dictionary form of a word. In some cases, it can be the same as the stem or the word, but in others it can be quite different from them.[1] For example, the Arabic surface form خاطب has the string خاطب as both a lemma and a stem, while the formتلتقيان *they meet* has the stem لتقي but the lemma التقى *he met*. For the present purposes, we have used the Farasa lemmatizer because it outperforms other tools such as Madamira (Pasha et al., 2014), Xerox Arabic morphological analysis and generation (Beesley, 1996), or Khoja's Stemmer (Khoja & Garside, 1999). Under lemma information, we provide measures for six variables, as illustrated in Fig. 7.

In Fig. 8, we display the output of a lemma search for the example word غاية *goal/purpose*.

Like the *Word Frequency* measure, the *Lemma Information* measure can also be queried using a list of words uploaded as a text file or using the *Constraints to Word* option to generate a list of stimuli with the desired properties.

Orthographic structure

Under this measure, two statistics are provided: the *number of letters* and whether there are *repeated letters* in a word. The number of letters returns the length of the input. In contrast, the statistics in *repeated letters* implement a Boolean search, returning 1 if the word has repeated letters (e.g., ممنوع 'mamnwE', *prohibited*) and 0 otherwise (e.g., مكتوب 'maktwb', *written*). Figure 9 displays the output of an orthographic structure search with the word ممنوع as an example.

The batch search using a list of words and the constraints to word search are also available for orthographic structure, affording the researcher an easy and fast way to compile their experimental materials.

Orthographic neighborhoods

This rubric subsumes the largest number of measures, totaling 20, augmented by two further measures, namely the bigram and trigram frequencies, as illustrated in Fig. 10.

Each of these measures is accompanied by a help menu laying out what it means, and each can be queried using a single search item or a list of items, as well as using the *Constraints to Word* option.

[1] There are different views about the stem in Arabic; for example, the same surface form (e.g., 'يكون yukawwin' *he forms*) is considered to have the stem 'kawwin' by Benmamoun (2003) but the stem 'kun' by Heath (2003). The same discord holds for natural language processing researchers (see Alshalabi et al., 2022).

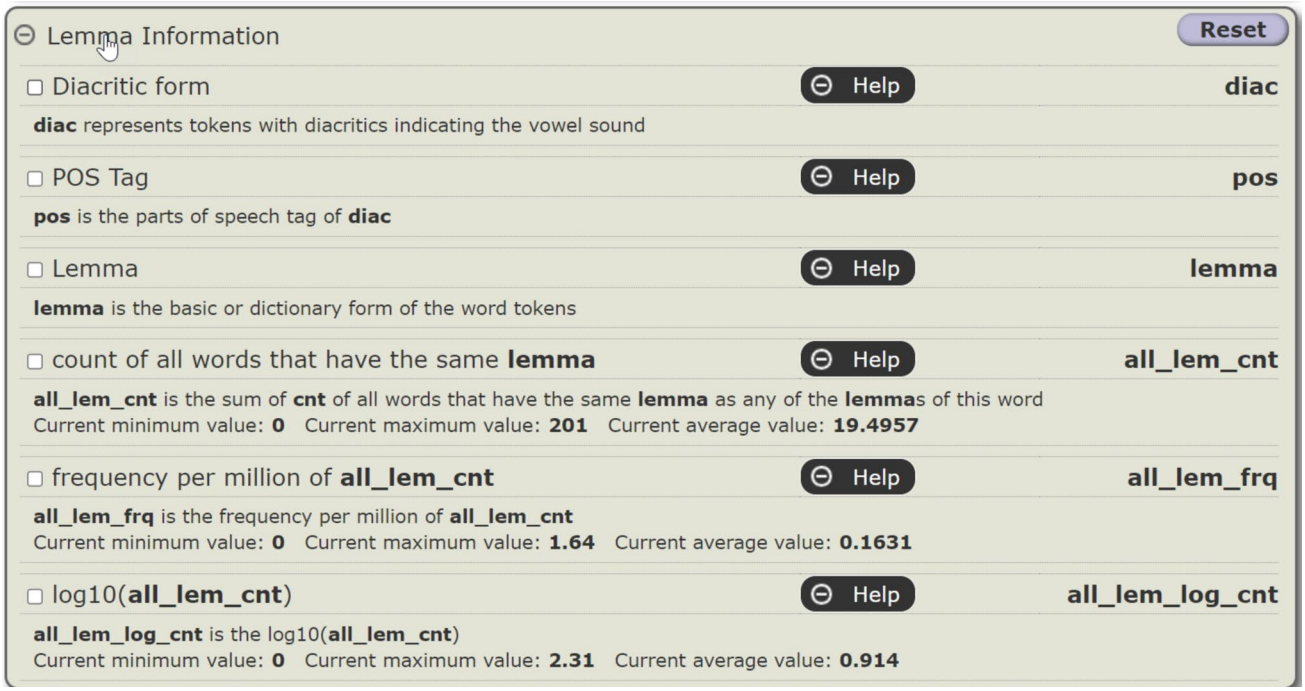


Fig. 7 Screenshot showing the six variables related to the *Lemma Information* measure with the *Help* menu displayed for the variable *Diacritic form* and *POS Tag*

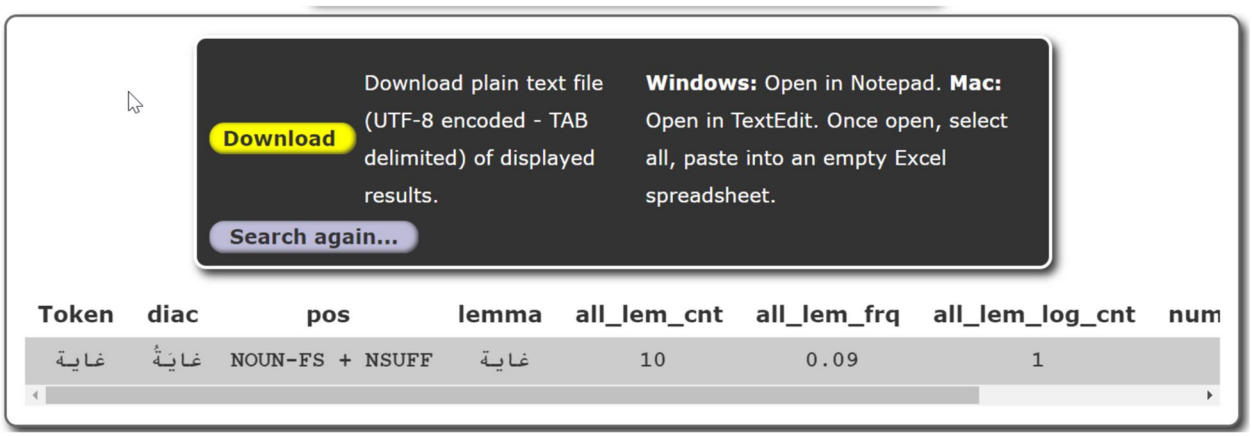


Fig. 8 Screenshot showing an example of a *Lemma Information* search for the word غاية *goal/purpose*

Phonological measures

SUBTLEX-AR is the first Arabic language database to provide statistics about the phonological domain in Arabic (excluding classical Arabic). The following two measures are covered:

Phonological structure

In linguistic parlance, phonological structure is a multifaceted concept that refers to different variables such as the phonotactic rules underlying the way in which speech sounds, or phonemes, are combined to create meaningful units, what syllables are grammatical in a given language, and how stress is assigned within a word. For the present purposes, we use the phrase phonological structure to refer to the number of phonemes a given word consists of. To compute this number, we counted the number of characters in the Farasa diacritized string written in Buckwalter transliteration. For instance, the Arabic form خرج *exit* is diacritized by Farasa as خَرَجَ, and its Buckwalter transliteration

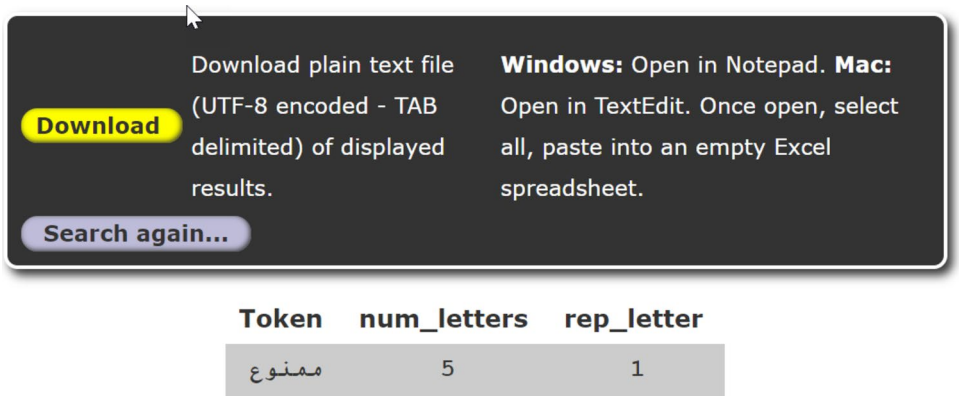


Fig. 9 Screenshot showing an example of an *Orthographic Structure* search for the word ممنوع *prohibited*

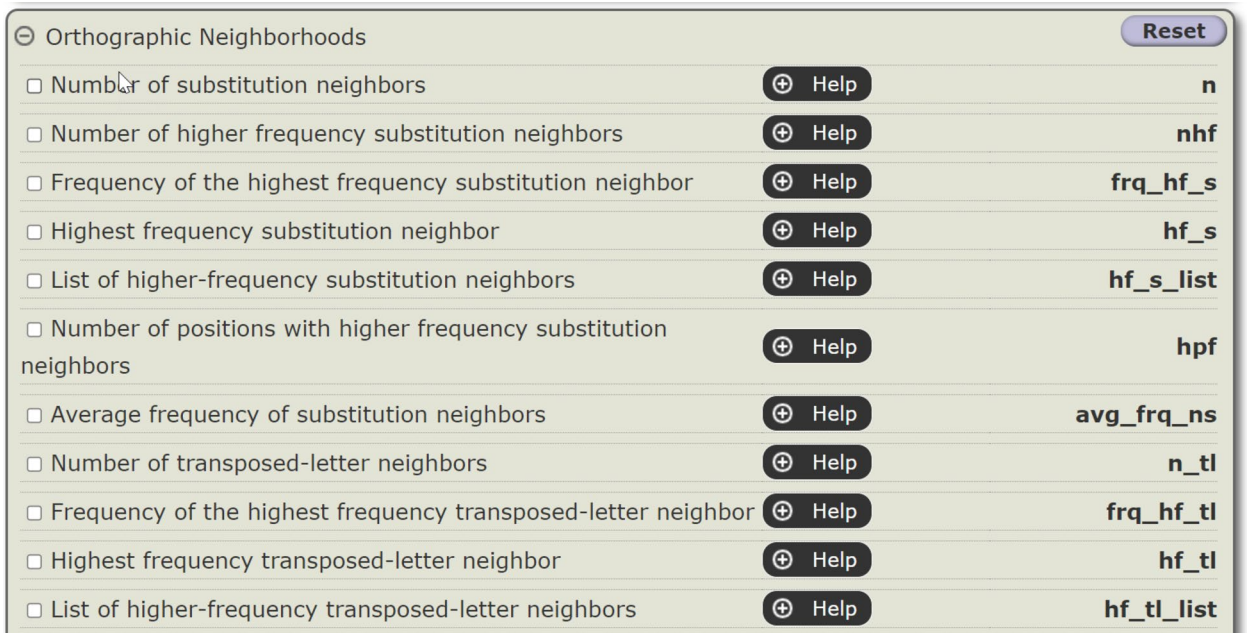


Fig. 10 Screenshot showing a sample of the measures available under the *Orthographic Neighborhoods* rubric

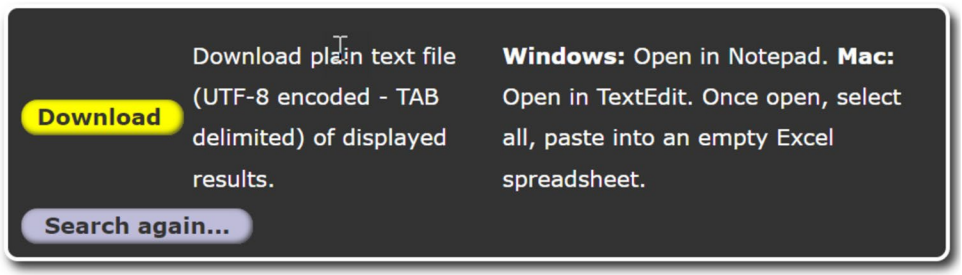
is 'xaraja' with six characters and accordingly six phonemes. Similarly, the form اليوم *today* is diacritized as اليَوْمَ transliterated as 'Alyawoma'[2] and yields a phoneme count of 7. As for the *number of syllables*, we subtracted the number of letters from the number of phonemes to obtain the syllable count for each word. For instance, the word معلومات, *information*, consists of 7 letters and 10 phonemes, 'maEolwmAt', and thus its syllable count is 3. An example query on phonological structure using a list of words is displayed in Fig. 11. Although this procedure is not ideal, it has a reasonable accuracy rate (more than 96%).

In Fig. 12, we display the results of a *Constraint to Words* search, setting the minimum and maximum number of phonemes to 3 and 10, respectively, and the minimum and maximum number of syllables to 3 and 6.

Phonological neighborhoods

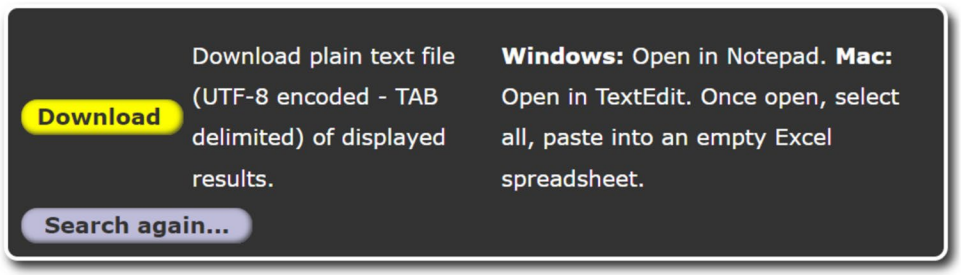
Phonological neighborhood refers to the number of words that differ in phonetic structure from another word based on a single phoneme that is substituted, deleted, or added (Luce & Pisoni, 1998). This rubric includes eight measures, each with its help menu, as shown in Fig. 13.

[2] The letter 'o' in the Buckwalter transliteration marks the absence of a vowel, it is called 'sukwn'—literally 'silence'.



Token	num_phon	num_syll
تقويم	8	3
جناح	6	2
خرج	6	3
شرب	6	3
قبيلة	8	3
كتاب	6	2
معلومات	10	3

Fig. 11 Screenshot showing an example of the output for the *Phonological Structure* rubric



Token	num_phon	num_syll
آباءهم	9	3
آباؤهم	9	3
آبائكم	9	3
آبائهم	9	3
آتيت	7	3
آثارهم	9	3
آثمة	7	3
آخذة	7	3

Fig. 12 Screenshot showing an example of the output of a *Constraint to Words*-based search on the *Phonological Structure* rubric with min–max set to 3–10 for the number of phonemes and 3–6 for the number of syllables

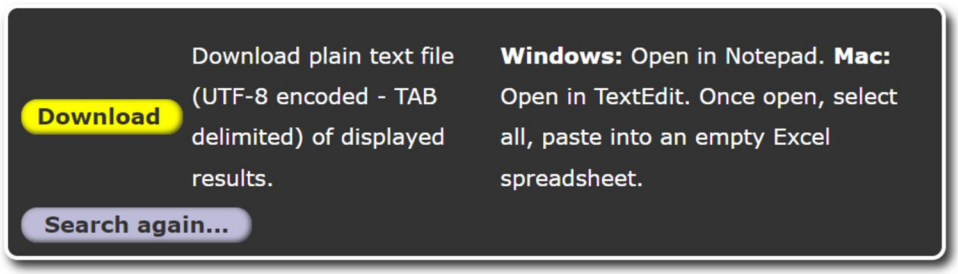
Like the previous rubrics, the measures under this one can be queried using a single word or a list of words to seek the phonological neighborhood characteristic of a preselected word or set of words. In addition, the *Constraints to Words* option allows the researcher to select word sets with specific properties. Figure 14 shows the output of *Constraints to words*-based search of the phonological neighborhood properties.

Morphological measures

This rubric of SUBTLEX-AR provides statistics about the type and token counts of the two main morphological components in Arabic, the root and the word pattern (hereafter WP), along with information about the stem. To identify the stem, root, and WP, we used Pyaramorph, a Python implementation of the Buckwalter Arabic morphological analyzer, which provides quick successive analyses of single words or short phrases (https://pypi.org/project/pyaramorph/). It accepts Unicode UTF-8 encoding as input and outputs a fully vowelled solution of all the possible readings of the word, part of speech (POS), and corresponding English glosses.

Where the stem is concerned, we extracted it from the POS results of Pyaramorph, which yields a stem after stripping off all affixes (e.g., يستمتعون 'ystmtEwn' *they enjoy themselves,* has the unpointed ستمتع 'stmtE' and the pointed stem سَتَمَتَّعْ 'satamat~aEo'). The frequency of the pointed stem thus obtained was calculated as the sum of frequencies of all word forms that share the same pointed stem, while the frequency of the unpointed stem is the sum of the frequencies of all its pointed forms. Turning to the root, we exploited the 'dictStems' file of Pyaramorph, an exhaustive lookup table that stores entries for all Arabic roots with their corresponding lemmas. We specifically matched the stem results obtained from the Pyaramorph POS tag segmentation against the root entry in the 'dictStems' table and seamlessly honed on the correct root for the word token at hand. There were a few instances where more than one root solution matched a particular stem. For example, the stem أمس '>ms', *yesterday*, matched the roots 'msw' *evening* and 'msy' *emaciate*. The largest number of root solutions was 4; hence, we provide all the solutions for each root. Finally, we established the WP as the residual of the stem once the root consonants have been stripped out and replaced by the letters 'f, E, l' as placeholders for the first, second, and third root letters, respectively. So, for example, the unpointed stem أسبوع, '>sbwE', *week*, is pointed as أُسْبُوعٌ '>usobwEN', and its root is identified as 'sbE', leading to the WP أُفْعُولٌ '>ufoEwlN', which

Fig. 13 Screenshot showing the eight measures available underneath the *phonological neighborhood* rubric



Token	NP	NPHF	frq_hfp	pf	pf_hf
آثم	9	5	841.25	3	2
آثمة	5	5	226.65	2	2
آرام	10	5	305.46	3	3
آسرة	9	9	450.89	3	3
آسفي	6	5	763.16	2	2
آفة	9	8	450.89	2	2
آكلها	8	5	163.5	2	2

Fig. 14 Screenshot showing a *Constraints to Words*-based phonological neighborhood search with min–max set to 5–10 for number of phonological neighbors and number of higher-frequency phonological neighbors, 100–1000 for frequency of the highest-frequency phonological neighbor and average frequency of phonological neighbors, and 1–3 for number of phonemes with phonological neighbors and number of phonemes with higher-frequency phonological neighbors

is the residual of the pointed stem with root consonants replaced by the place holders 'f, E, l'.

Stem, root, and WP details

This measure provides the statistics for the stem, the root, and the WP. It subsumes 28 pieces of information, including the frequency statistics for pointed and unpointed stems, the type and token frequency for different roots, and WPs. Figure 15 displays an example search for this rubric using the search item أمس '>ms', *yesterday*.

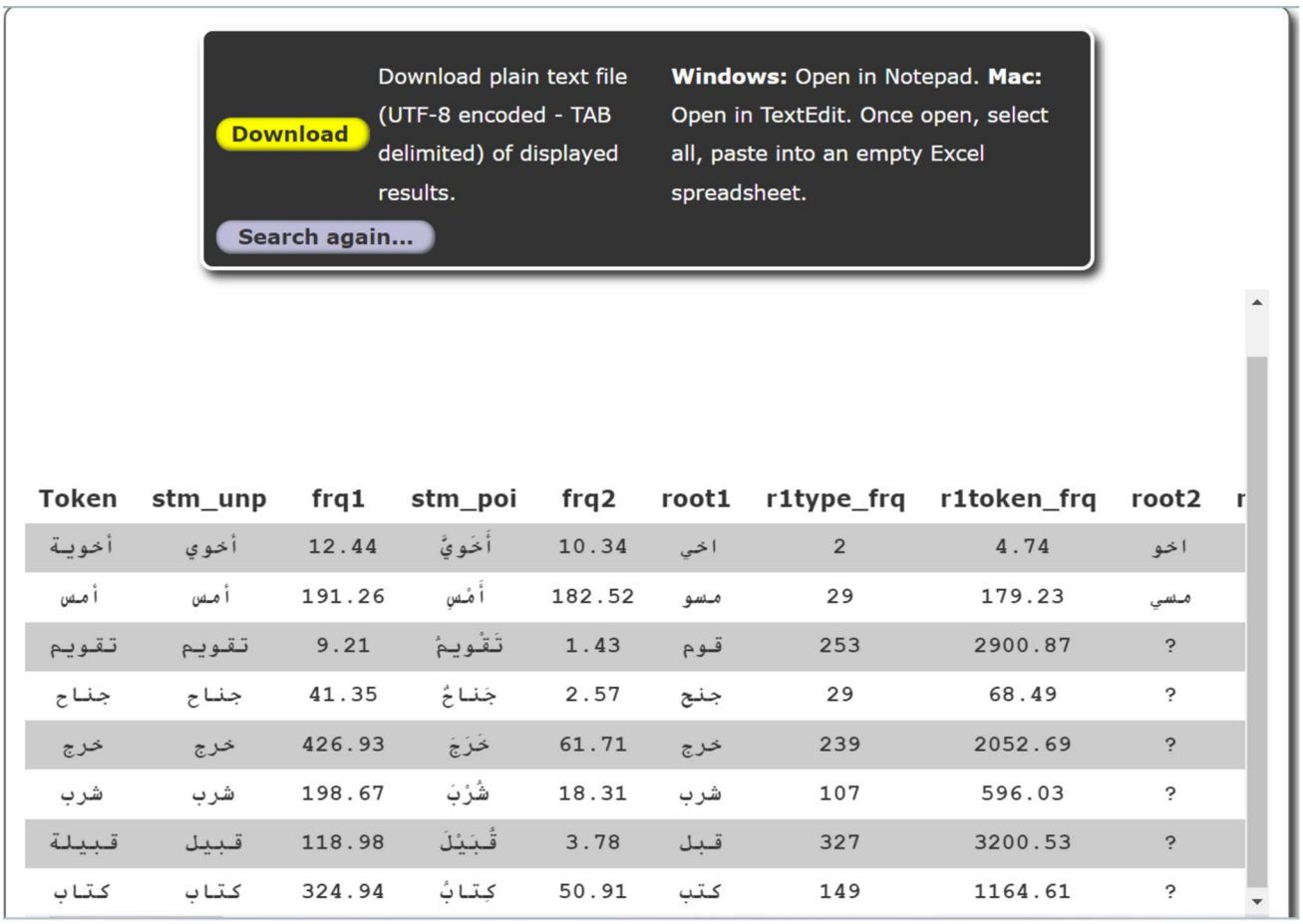
Diacritic information

This is one of the strongest and most original aspects of SUBTLEX-AR. It provides the different diacritized lemmas for each input string and its frequency, as shown in Fig. 16. Given the highly ambiguous nature of Arabic orthography, this functionality provides frequency counts not only for different inflectional variants of the same word (e.g., حَمَّامٌ 'Ham~AmN' = *a bath, nominative*; 'Ham~Ama' = *a bath, accusative*; 'Ham~Ami' = *a bath, genitive*) but also for heterophonic homographic variants (e.g., حَمّامٌ 'Ham~AmN' = *a bath, nominative;* حَمامٌ 'HamAmN' = *pigeons, nominative).*

Semantic measures

SUBTLEX-AR is unique even among other SUBTLEX databases, as it is the first to provide a set of quantitative semantic measures, enabling users to explore relationships between words in terms of meaning. These semantic relationships are modeled using Word2Vec, a computational method that captures word relationships based on their contextual co-occurrences.

Specifically, we trained a continuous bag of words (CBOW) model, a type of Word2Vec model, on the entire corpus of Arabic subtitles. The CBOW model works by predicting a target word based on the surrounding words, which allows it to learn the patterns of word co-occurrence. The assumption is that words that frequently appear in similar contexts will have similar meanings. For example, if the



Token	stm_unp	frq1	stm_poi	frq2	root1	r1type_frq	r1token_frq	root2	r
أخوية	أخوي	12.44	أَخَويٌّ	10.34	اخي	2	4.74	اخو	
أمس	أمس	191.26	أَمْسِ	182.52	مسو	29	179.23	مسي	
تقويم	تقويم	9.21	تَقْويمُ	1.43	قوم	253	2900.87	?	
جناح	جناح	41.35	جَناحُ	2.57	جنح	29	68.49	?	
خرج	خرج	426.93	خَرَجَ	61.71	خرج	239	2052.69	?	
شرب	شرب	198.67	شُرْبَ	18.31	شرب	107	596.03	?	
قبيلة	قبيل	118.98	قُبَيْلَ	3.78	قبل	327	3200.53	?	
كتاب	كتاب	324.94	كِتابُ	50.91	كتب	149	1164.61	?	

Fig. 15 Screenshot showing part of the output of a *Stem Root Pattern Details* search using a list of words. Note: The user can slide to the right to visualize the rest of the measures. The question mark underneath root2 means that it has a single root solution

Token	wp4token_frq1	diac1	diac_frq1	diac2	diac_frq2	diac3	diac_frq3	diac4
حمام	0	حَمّامٌ	615	حَمّامَ	492	حَمّامِ	1475	حَمّامُ
خال	0	خال	567	خالِ	1059	خالَ	201	خالُ
سحابة	0	سَحابَةً	114	سَحابَةُ	127	سَحابَةٌ	120	سَحابَةِ
سلم	0	سَلَّمَ	1383	سَلَّمِ	80	سُلَّمِ	179	سَلَّمَ
صاحب	0	صاحِبُ	4734	صاحِبِ	1551	صاحِبَ	4210	صاحِبْ
علم	0	عِلْمِ	3205	عِلْمِ	3290	عِلْمَ	5522	عِلْمٌ
كتاب	0	كِتابِ	3598	كِتابَ	854	كِتابُ	2548	كِتابٌ

Fig. 16 Screenshot showing part of the output of a *Diacritic Information* search using a list of words. Note the different readings and their associated frequency counts

words ‘book’ and ‘library’ often occur together in sentences, the model will learn that they are semantically related.

The model encodes each word as a 300-dimensional vector, where each dimension represents a feature of the word’s meaning learned from the text. While these vectors are not directly interpretable by humans, they effectively position words in a semantic space where words with similar meanings are positioned close to one another.

To compute the semantic similarity between two words, we calculate the cosine similarity between their vectors. This

method measures the angle between the two-word vectors: If the vectors are pointing in almost the same direction (which indicates that the words have similar meanings), the cosine of the angle will be close to 1. Conversely, if the vectors are orthogonal (indicating that the words are unrelated), the cosine will approach 0. Thus, cosine similarity provides a score between 0 and 1, where values closer to 1 indicate higher semantic similarity.

This approach in SUBTLEX-AR allows researchers to quantitatively assess the semantic relationships between words, offering a powerful tool for fine-grained analysis of language data.

Most similar words

This rubric lets users retrieve the ten nearest neighbors and their cosine similarity values. Figure 17 displays the ten nearest neighbors of the word عميل, 'Emyl', a homonym with two unrelated meanings, one pertaining to *spying* and the other to *being a customer*. Remarkably, the output captures this homonymic aspect of the word, yielding words related to its two meanings, with such examples as مخبر, 'muxobir', *informant,* and زبون, 'zabwn', *customer/client*.

Word pair similarity

This functionality allows users to input pairs of words, either one word pair at a time or a list of word pairs, and obtain their cosine similarity. As with other searches, the input can be either in Arabic or Buckwalter transliteration. Figure 18 below shows an example of a search with a list of nine word pairs.

Token	nn_words	Similarity
عميل	مخبر	0.74
عميل	موكل	0.69
عميل	جاسوس	0.69
عميل	سجين	0.69
عميل	مستثمر	0.68
عميل	زبون	0.68
عميل	موظف	0.67
عميل	محقق	0.67
عميل	وعميل	0.66
عميل	ضابط	0.66

Fig. 17 Screenshot showing the output of a *Most Similar Words* search using the word عميل, Emyl, *spy/customer* as a search term

Word 1	Word 2	Similarity
جميل	رائع	0.835
حار	بارد	0.754
دقيق	رشيق	0.45
دقيق	عجين	0.211
طبق	وجبة	0.442
أخرج	تخرج	0.668
أخرج	خراج	0.019
أخرج	أهدر	0.291
أخرج	بدلة	0.009

Fig. 18 Screenshot showing the output of a *Word Pair Similarity* search

This word sample illustrates a range of relations, including synonymy (رائع-جميل 'jmyl'-' rA}E', *beautiful-wonderful*), antonymy (بارد-حار 'Har'-'bArd', hot-old), polysemy (عجين-دقيق, 'ddyq'-'Ejyn', *flour-dough*, and رشيق-دقيق 'dqyq'-'r$yq' *graceful-delicate*), and metonymy (وجبة-طبق 'Tbq'-'wjbp' *meal-dish*). Because morphemes are minimal meaningful units that cannot be divided any further, and because they are prominent elements of lexical processing, the sample also includes a pair of words that share a root and a transparent semantic relationship (تخرج-أخرج '>xrj'-'txrj', *get out-pull out*), a pair that shares a root and an opaque semantic relationship (خراج-أخرج '>xrj'-'xrAj', *get out-tribute*), and a pair that shares a WP (أهدر-أخرج '>xrj'-'>hdr', *get out-waste*). The final pair comprises two random words (بدلة-أخرج '>xrj'-'bdlp', *get out-suit*) that are not related along any dimensions at all.

Figure 18 suggests that the SUBTLEX-AR word embeddings capture reasonably well the different kinds of relationships, with synonyms, antonyms, metonyms, and words sharing a root and a transparent semantic relationship having the highest similarity scores. In contrast, words that share a root and a nontransparent semantic relationship have a similarity score that is inferior to that

assigned to words related by a WP but greater than that assigned to a random word pair.

SUBTLEX-AR validation

Although the primary focus of this article is to report the origin and processing methods of the word frequency data in SUBTLEX-AR, we present in this section a validation of the different psycholinguistic measures it provides. Specifically, we report a lexical decision experiment involving 1000 words[3] and 1000 nonwords and conduct multiple regression and correlation analyses on latency and accuracy data to determine the percentage of variance accounted for by the different metrics. For the orthographic measures, we compared SUBTLEX to ARALEX. For the morphological and phonological measures, we report the amount of variance in response latencies and accuracies accounted for by these variables. This is because the morphological variables, namely root and WP type and token frequencies, are very similar in SUBTLEX-AR and ARALEX, and the phonological variables are not available in ARALEX or any other database. Finally, we evaluate the semantic measures against semantic judgments made by human participants and against the Word2vec embeddings acquired from the publicly available fasttext.cc (https://fasttext.cc/docs/en/crawl-vectors.html) developed by Bojanowski et al. (2017).

Lexical decision experiment

The lexical decision task (LDT) is a ubiquitously used word recognition task in which participants are required to determine whether a string of letters is a word or not. We chose to use it for a number of reasons. First, this task allowed us to maintain comparability with how previous SUBTLEX databases have been validated (e.g., Brysbaert et al., 2011; Brysbaert & New, 2009; Cuetos et al., 2011; Dimitropoulou et al., 2010; Keuleers et al., 2010; van Heuven et al., 2014a, 2014b). Second, this task has been repeatedly shown to be highly sensitive to benchmark phenomena in the literature on word identification, such as word frequency effects (Balota et al., 2004) and morphological effects (e.g., Marslen-Wilson et al., 1994; Boudelaa et al., 2023). Finally, this task is easy to design for the experimenter and to carry out for the participants in the laboratory and online.

[3] The words used for lexical decision were selected before completing the development of SUBTLEX-AR, and to our surprise, upon completion of the database, it turned out that 70 words were not available in SUBTLEX-AR, and a further two were repeated, and so the statistical analyses we report are based on 928 word items.

Method

Participants One hundred and ten student volunteers took part in the experiment. All of them were native speakers of Arabic with no history of reading problems. The experiment was approved by the Research Ethics Committee of the United Arab Emirates University following the Declaration of Helsinki, and participants gave their consent before participating.

Materials We used a total of 1000 words varying in length from 2 to 8 letters ($M = 4.09$, $SD = 0.98$). Their OLD20 ranged from 1 to 2.8 ($M = 1.70$, $SD = 0.26$). For the purposes of the lexical decision task, 1000 nonwords were included. These were closely matched in terms of length ($M = 4.09$, SD = 0.98) and OLD20 ($M = 1.93$, SD = 0.24) to the words. The nonwords were formed by changing or adding 1 to 2 root letters to an existing word (e.g., كسوف 'kswf' *eclipse*) to create a nonword (e.g., جسوف 'jswf') that consisted of a nonexisting root (e.g., 'jsf') and an existing Word Pattern (e.g., 'fEwl').

Design The words and nonwords were assigned randomly to four blocks of 500 stimuli (250 words and 250 nonwords). Participants were allowed three self-paced breaks. Stimuli were presented centrally on a computer screen in black letters against a white background (Traditional Arabic, 32 pt. regular). A trial started with the presentation of a fixation symbol '+' at the center of the screen. Participants were asked to fixate on the '+' sign as soon as it appeared. After 500 ms, the fixation point disappeared, and a target item was presented. The stimulus stayed on the screen until the participant responded or 2 s had elapsed. The interstimulus interval was 350 ms. Participants used their dominant hand to press the forward slash key '/' for a word response and their other hand to press the 'Z' key for a nonword response. Auditory feedback was provided for correct responses (high-pitched tone) and incorrect responses (low-pitched tone). Stimulus presentation was randomized for each participant. The experiment was programmed using SuperLab software (Cedrus, Phoenix, AZ, USA). The experiment lasted about 15–20 min and began with 30 training trials.

Results and discussion

The experimental trials comprised 62,000 data points; half were word responses, and half were nonword responses. After discarding the latencies for word items that were inadvertently included in the experiment but did not have the relevant measures in SUBTLEX-AR, 28,767 remained. We inspected the Q-Q plot and pruned out reaction times (RTs)

above 1900 ms or below 100 ms as outlying data points. There were 3.46% errors and 0.79 outliers for the word data, while for the nonwords (31,000 data points), the percentages of errors and outliers were 5.81 and 0.97, respectively. The average RT for words was 632.42 ms (SD: 225.68), and for nonwords, 737.06 ms (SD: 282.67).

We fitted a series of linear mixed-effects (LMER) models[4] estimated using maximum likelihood (ML) and the optimx optimizer to predict the logarithm of the RTs (logrt) with the variables in a given domain and a series of logit models on the accuracy data (Bates et al., 2015; Kuznetsova et al., 2017). There is no satisfactory method to determine the proportion of variance accounted for by each predictor in a linear mixed-effects model, and those that exist only provide a global R^2 for all fixed effects taken together (Nakagawa & Schielzeth, 2013). Accordingly, we conducted multiple Pearson's product–moment correlations, which we then squared and multiplied by 100 to obtain an R^2, a unitless summary index that can be objectively used for cross-studies comparisons (Nakagawa & Cuthill, 2007). The complete list of materials and analysis scripts are freely accessible on OSF: https://osf.io/spb8c/

Orthographic and morphological variables

Latency and accuracy data

We began by analyzing the predictors of the orthographic and morphological domains together because they are largely common to both SUBTLEX-AR and ARALEX and thus allow us to compare the two databases. Accordingly, we fitted an LMER model that included word frequency, pointed stem frequency, unpointed stem frequency, and OLD20 for both SUBTLEX-AR and ARALEX (Yarkoni et al., 2008). Where the morphological predictors are concerned, the model also included root and WP type and token frequencies for SUBTLEX-AR only because the type measures (i.e., root type and WP type) are identical across SUBTLEX-AR and ARALEX, and the token frequencies (i.e., root token and WP token) were highly correlated. The predictors lemma type frequency (all_lem_cnt) and lemma token frequency (all_lem_frq) were also included only for SUBTLEX-AR because ARALEX does not provide these measures. Finally, the model included target and block as random effects.[5]

The results revealed that the total variance explained by the fixed effects alone was 34.41%. The effect of word frequency was statistically significant for both SUBTLEX-AR ($F(1, 914.91) = 142.42$, $p < 0.001$) and ARALEX ($F(1, 913.07) = 157.91$, $p < 0.001$), as was the effect of OLD20 for SUBTLEX-AR ($F(1, 913.38) = 61.64$, $p < 0.001$) and ARALEX ($F(1, 914.07) = 136.73$, $p < 0.001$). In contrast, unpointed stem frequency revealed a significant effect for ARALEX ($F(1, 913.26) = 19.03$, $p < 0.001$) but no effects for SUBTLEX-AR ($p = 0.22$). The effects of root type frequency ($F(1, 914.31) = 16.40$, $p < 0.001$), WP type frequency ($F(1, 913.62) = 7.56$, $p < 0.01$), and lemma type frequency ($F(1, 912.86) = 55.24$, $p < 0.001$) were all statistically reliable.

To determine the amount of variance accounted for by each individual variable that had a significant effect in the LMER model, we ran a series of Pearson's product–moment correlations and computed the R^2 values for each correlation. Then we compared the reliability of the difference between the amount of variance accounted for by each of the SUBTLEX-AR and ARALEX frequency measures using the online version of the cocor R package available at http://comparingcorrelations.org/, setting the parameters to two overlapping correlations based on dependent groups with sample size $n = 27{,}617$, alpha = 0.05, and CI = 0.95. The results of these analyses are displayed in the last column of Table 1.

Table 1 reveals that SUBTLEX-AR outperforms ARALEX on all variables except the morphological ones (i.e., root and WP measures), which are the same for the two databases. The better SUBTLEX-AR performance resonates with previous research in this area, wherein SUBTLEX databases often outperform the traditional databases that are based on newspaper articles and other literary genres (e.g.,

Table 1 Percentages of variance in reaction times (RT) explained by different orthographic and morphological measures in SUBTLEX-AR and ARALEX and the reliability of the difference between the two databases measured by Pearson and Filon's *z*

Predictor	RT% Variance		
	ARALEX	SUBTLEX-AR	Pearson and Filon's Z
Word frq.	9.04	19.85	z = −19.243, $p = 0.000$
OLD20	9.91	15.96	z = 10.738, $p = 0.000$
Unpointed stem frq.	9.83	11.24	z = −3.0180 $p = 0.0025$
Root type frq.	4.81	4.81	z = 0.0000, $p = 1.000$
WP type frq.	0.03	0.03	z = 0.0000, $p = 1.000$
Lemma type frq.	n/a	15.9	n/a

Root type frq and WP type frq are identical in the two databases, while lemma type frq is available only in SUBTLEX-AR

[4] We tested models with the log10-transformed frequencies for the different measures adding 1 and applying a polynome (degree 3), but these manipulations did not improve model fit.

[5] We removed Participants from the model because it gave rise to singularity issues.

Brysbaert & New, 2009; Gimenes & New, 2016; Keuleers et al., 2010; Soares et al., 2015). The six variables in Table 1 fall neatly into two sets. The first set consists of word frequency, unpointed stem frequency, and root type frequency and exerts a facilitative effect on RT in the sense that the items with the higher values for these measures have faster latencies and vice versa. The second set consists of WP type frequency, OLD20, and lemma type frequency and has an inhibitory effect with higher values in any of these measures, leading to slower RTs.

Turning to the accuracy data, we fitted a logit model with the same structure as the LMER model described above, except that the dependent variable was the error rate. In this model, the fixed effects alone accounted for 34% of the variance. The only effects that were reliable were word frequency ($B = 0.002$, $Z = 5.469$, $p < 0.001$), pointed stem frequency ($B = 0.001$, $Z = 3.339$, $p < 0.001$), and OLD20 ($B = -0.632$, $Z = 2.755$, $p < 0.005$) for SUBTLEX-AR, and all of these had an inhibitory effect on response accuracy. A point–biserial correlation coefficient analysis suggests that the word frequency and pointed stem frequency, respectively, account for 11.04% and 10.22% of the variance in the response accuracy. In comparison, OLD20 accounted only for a negligible 0.09%. Finally, none of the homologous ARALEX variables was significant in the logit model, and so were not analyzed any further.

Phonological variables

Latency and accuracy data

In the model,[6] we fitted logrt ~ num_phon + num_syll + NP + NPHF + frq_hfp + pf + pf_hf, with target and block as random effects. Of the total variance, 49.97% was explained by the fixed effects alone. In addition, the results revealed the effects of num_phon ($F(1, 913.76) = 68.94$. $p < 0.001$), num_syll ($F(1, 914.78) = 90.15$, $p < 0.001$), NP ($F(1, 913.58) = 192.46$, $p < 0.001$), NPHF ($F(1, 914.94) = 771.63$, $p < 0.001$), pf ($F(1, 913.61) = 9.71$, $p < 0.002$), and pf_hf ($F(1, 914.17) = 125.08$, $p < 0.000$) to be statistically significant. Additional Pearson correlations revealed that the amount of variance explained by each variable was as follows: num_phon: 2.85%, num_syll: 0.04%, NP: 14.03%, NPHF: 9.46%, pf: 1.89%, and pf_hf: 9.74%. The effects of num_syll, NP, frq_hfp, and pf on the RTs were facilitative, while those of num_phon, NPHF, and pf_hf were inhibitory.

[6] Number of phonemes (num_phon), number of syllables (num_syll), number of phonological neighbors (np), number of higher-frequency phonological neighbors (NPHF), frequency of the highest frequency phonological neighbor (frq_hfp), number of phonemes with phonological neighbors (pf) and number of phonemes with higher frequency phonological neighbors (pf_hf).

Finally, a logit model with the same structure as the LMER model used for the latency data revealed NP to be the only statistically reliable effect ($B = 0.013$, $Z = 2.132$, $p = 0.033$) with an inhibitory effect and accounting for 0.64% of the variance in the accuracy data.

Semantic variables

To assess the validity of the semantic representation developed for SUBTLEX-AR, we chose 588 word pairs for which we had previously gathered semantic relatedness ratings for a different study. In that study, 15 participants were asked to rate pairs of items on a 9-point scale of semantic relatedness, with 9 representing 'very related in meaning' and 1 representing 'very unrelated in meaning'. The same 588 word pairs also had semantic representations in the Word-2vec embeddings database fasttext.cc (Bojanowski et al., 2017) and in SUBTLEX-AR. Figure 19 below displays the correlation between SUBTLEX-AR and both human participants' ratings and the CC.300 and the correlation between the latter two.

It is worth noting that the correlation between SUBTLEX-AR and human participants' ratings ($r = 0.76$, $p < 0.001$) is nearly three times as high as that between human ratings and the CC.300 ($r = 0.26$, $p < 0.001$), clearly demonstrating that the SUBTLEX-AR word embeddings outperform the vectors from CC.300 and mirror much more closely the semantic representations developed by native speakers.

General discussion

The results of our validation of the different measures of the subtitles database in the Arabic SUBTLEX-AR (some against the ARALEX database on newspaper articles) clearly show that this new database promises to provide a valuable resource for researchers across many fields. For the three measures of word frequency, OLD20, and unpointed stem frequency, SUBTLEX-AR accounted for 47% of the variance of the RT data, while ARALEX accounted for 28.78%. For the accuracy data, the SUBTLEX-AR word frequency, OLD20, and pointed stem frequency were significant, respectively accounting for 33%, 31%, and 0.2% of the variance. For the morphological variables, root type frequency appears to have a consistent modulatory impact, accounting for 4.81% of the variance in response latencies with no effects on the accuracy response. Lemma type and lemma token frequencies seem identical, and the amount of variance they each account for in RT is 15.9%. In comparison, the amount of variance they account for in the accuracy data is a nonsignificant at 0.2%.

Thus, the strongest orthographic predictors are word frequency, OLD20, and lemma type (or token) frequency.

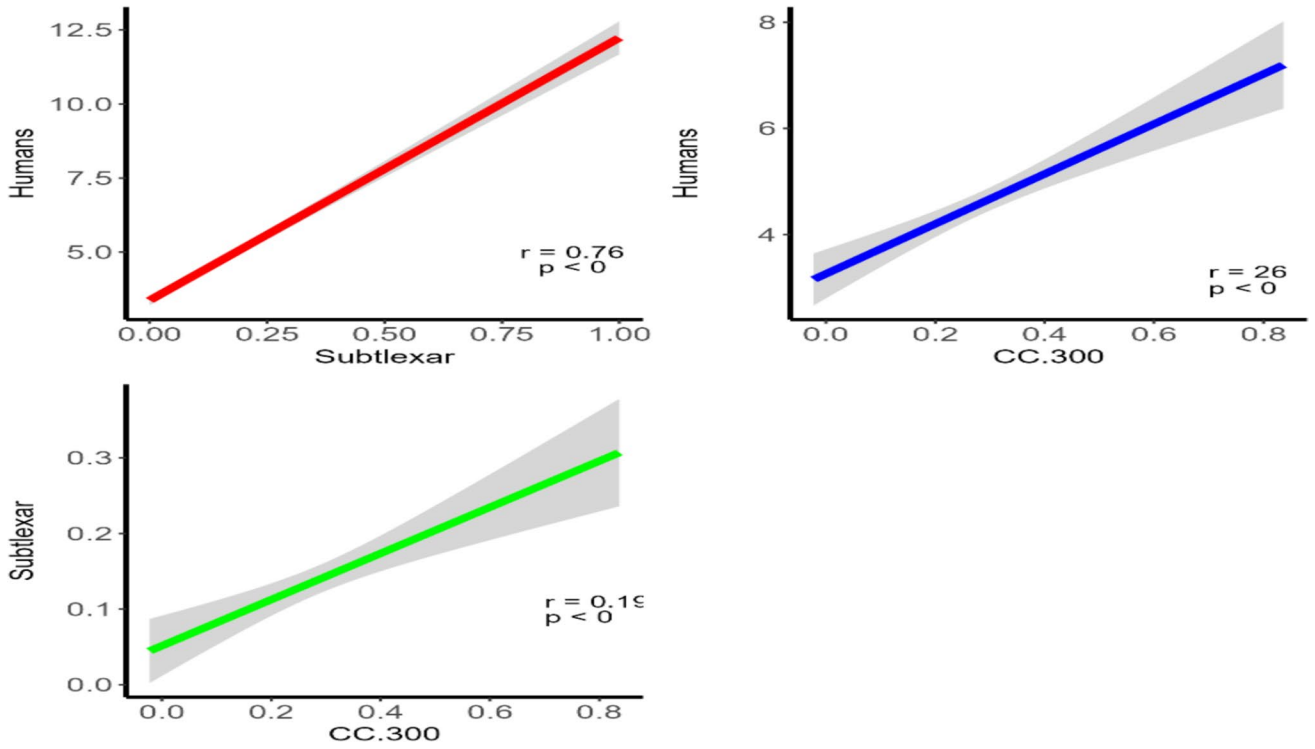


Fig. 19 Pairwise correlations between SUBTLEX-AR, human ratings, and CC.300 embeddings

In contrast, the strongest morphological measures are root type frequency and unpointed stem frequency. It is interesting—and perhaps surprising—that only two out of six morphological variables (i.e., root type and token, WP type and token, and stem pointed and unpointed) turn out to modulate response times in a Semitic language where morphological effects have consistently been claimed to be pervasive (Barkai, 1980; Boudelaa & Marslen-Wilson, 2001, 2004a; 2004b; 2011; 2015; Boudelaa et al., 2023; 2024; Frost et al., 1997, 2000a, 2000b, 2005; Prunet et al., 2000; Perea et al., 2011, 2014, 2010). One reason for this may be that the correlation is fairly high between root type and token frequencies ($r(27615) = .35$, $p < .001$), on the one hand, and between WP type and WP token frequencies on the other ($r(27615) = .23$, $p < .001$). Further research is needed to elucidate this issue. Finally, the phonological predictors account for 49.97% of the variance in the latency data and 35% in the accuracy data. The most prominent phonological variables are NP, followed by NPHF and pf_hf, collectively accounting for 33.23% of the variance in RTs.

Taken together, these results suggest that researchers can safely limit their efforts to the subset of frequency measures that we mentioned here and that play a major modulatory role in the response latencies. For researchers interested in orthographic processing in Arabic, it is arguably enough to control for word frequency, OLD20, and lemma type (or token) frequency. Researchers interested in the morphological domain can limit their stimulus-matching efforts to root type and unpointed stem frequencies. Where the phonological domain is concerned, we believe that matching stimulus materials on NP, NPHF, and pf_hf will suffice to construct a well-controlled set of items.

A second take-home message from the validation of the orthographic, morphological, and phonological domains in this research is that in Arabic, as in Indo-European languages (e.g., Dutch, English, French, German, Greek, Spanish, and Portuguese) and Sino-Tibetan languages (e.g., Chinese), frequency measures derived from movie subtitles are effective predictors of the word recognition processes in young adult college students, explaining a higher percentage of the variance in the word latencies than a database derived from newspaper articles.

A final aspect of our validation exercise that makes SUBTLEX-AR unique even among fellow SUBTLEX databases is the provision of word embeddings that are reliable and agree with participants' ratings, suggesting that these embeddings can go a long way in helping researchers select objectively controlled experimental items to investigate semantic processing. The SUBTLEX-AR word embeddings can be supplied to researchers upon request and can be used to run simulations (e.g., Armstrong & Plaut, 2016) and predict neural activity (e.g., Mitchell et al., 2008). As such, SUBTLEX-AR will find a niche among researchers interested in visual word recognition and reading in Arabic as a first language

and as a second language, and will be a welcome addition to recent databases such as MECO (Siegelman et al., 2022; see also Kuperman et al., 2022).

Conclusions

The current study presented a new lexical frequency measure for Arabic based on subtitles of films and TV series. SUBTLEX-AR explains more variance (18.27% more) in the latency data than the written-word frequency measures obtained from ARALEX, which was based on newspaper articles. Additionally, SUBTLEX-AR accounts for 37% of the variance in the accuracy of data pooling over orthographic, morphological, and phonological variables. This outcome is entirely in line with results from similar SUBTLEX databases found with a range of languages including English (both British and American), Greek, Chinese, Dutch, German, Spanish, or Portuguese. Compared with ARALEX, SUBTLEX-AR frequencies represent a notable improvement in explained variance in RTs, thus constituting a valuable resource for cognitive studies based on verbal materials. Although we believe that when it comes to stimulus selection, the lexical information contained in ARALEX still has a role to play, the breadth and reliability of the SUBTLEX-AR measures should be preferred over those of ARALEX. Users of SUBTLEX-AR will notice that the version we are sharing is a beta version. This is because we believe that despite the clear advantages of SUBTLEX-AR in its current form, we intend to release an updated version of the database. A first aspect to improve in future releases of SUBTLEX-AR will be the lack of statistics pertaining to contextual diversity (CD), a measure of the number of passages (documents/movies) in the corpus containing a given word. CD has been claimed to be better than word frequency in predicting RT (Adelman et al., 2006; Perea et al., 2013), and it will be interesting to determine whether, indeed, it accounts for more variance than word frequency in Arabic as it does in other languages (e.g., English: Brysbaert & New, 2009; Chinese: Cai & Brysbaert, 2010; Greek: Dimitropoulou et al., 2010; Dutch: Keuleers et al., 2010; Portuguese: Soares et al., 2015). A further area for improvement is the phonological measures, which currently rely on approximate diacritized Arabic values rather than precise phonetic transcriptions. Additionally, the database could potentially expand to cover phrase- and sentence-level data. Brysbaert et al. (2024) advocate for artificial intelligence (AI)-generated familiarity estimates to capture language knowledge beyond individual words, extending to phrases and multiword expressions (MWEs). For SUBTLEX-AR, a phrase-level expansion seems feasible and would enhance utility despite possible interface challenges, whereas a sentence-level expansion may prove overly complex at this stage. Finally, we also encourage users to share feedback on other aspects for enhancement, and we are committed to considering all suggestions for updates and improvements. SUBTLEX-AR is freely available for research purposes (https://subtlexar.uaeu.ac.ae/).

Authors' contributions MC, MP, and SB developed the study concept and contributed to the study design. Data collection and interface development was performed by NJ under the supervision of SB. Data analysis and interpretation were carried out by SB, who drafted the initial manuscript, and all authors provided substantial revisions. All authors approved the final version of the manuscript for submission.

Funding This research was funded by a UAEU grant (G00003452) awarded to Sami Boudelaa, Manuel Carreiras, and Manuel Perea. Additional support was provided through a UAEU CHSS grant (G00003887) to Sami Boudelaa, as well as contributions from the Dean's Office at CHSS-UAEU. This work was also supported by the Department of Education, Culture, Universities, and Employment of the Valencian Government through grant CIAICO/2021/172 awarded to Manuel Perea.

Availability of data and materials The data and materials supporting the findings of this study are openly available at https://osf.io/spb8c/.

Code availability The code used for analysis and supporting the findings of this study is openly available at https://osf.io/spb8c/.

Declarations

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to participate All participants provided informed consent prior to their involvement in the study. The nature, purpose, and procedures of the research were explained in detail, ensuring participants understood their rights, including the right to withdraw at any time without penalty. Participants were assured of the confidentiality of their data and its use solely for research purposes.

Consent for publication All participants provided explicit consent for the publication of anonymized data and findings derived from this study.

Conflicts of interest/Competing interests All authors declare no competing financial or non-financial interests.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17*, 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Alshalabi, H., Tiun, S., Omar, N., Al-Aswadi, F. N., & Alezabi, K. A. (2022). Arabic light-based stemmer using new rules. *Journal*

of King Saud University - Computer and Information Sciences, 34(9), 6635–6642. https://doi.org/10.1016/j.jksuci.2021.08.017

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience, 31*, 940–996. https://doi.org/10.1080/23273798.2016.1171366

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of experimental psychology. General, 133*(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Barkai, M. (1980). Aphasic evidence for lexical and phonological representations. *Afroasiatic Linguistics, 7*(6), 163–187.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beesley, K. R. (1996). Arabic finite-state morphological analysis and generation. *Proceedings of the 16th conference on Computational linguistics* (vol. 1, pp. 89–94). Association for Computational Linguistics. https://doi.org/10.3115/992628.992647

Benmamoun, E. (2003). The role of the imperfective template in Arabic morphology. In J. Shimron (Ed.), *Language processing and acquisition in languages of Semitic, root-based, morphology* (pp. 99–114). Amsterdam: John Benjamins.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Boudelaa, S., Boujraf, S., Belahcen, F., Ben Zagmout, M., & Farooqui, A. (2023). Impaired morphological processing: Insights from multiple sclerosis. *Language, Cognition and Neuroscience, 38*(9), 1237–250. https://doi.org/10.1080/23273798.2023.2226267

Boudelaa, S., & Marslen-Wilson, W. D. (2001). Morphological units in the Arabic mental lexicon. *Cognition, 81*(1), 65–92. https://doi.org/10.1016/s0010-0277(01)00119-6

Boudelaa, S., & Marslen-Wilson, W. D. (2004a). Abstract morphemes and lexical representation: The CV-Skeleton in Arabic. *Cognition, 92*(3), 271–303. https://doi.org/10.1016/j.cognition.2003.08.003

Boudelaa, S., & Marslen-Wilson, W. D. (2004b). Allomorphic variation in Arabic: Implications for lexical processing and representation. *Brain and Language, 90*(1–3), 106–116. https://doi.org/10.1016/S0093-934X(03)00424-3

Boudelaa, S., & Marslen-Wilson, W. D. (2010). ARALEX: A lexical database for modern standard Arabic. *Behavior Research Methods, 42*, 481–487. https://doi.org/10.3758/BRM.42.2.481

Boudelaa, S., & Marslen-Wilson, W. D. (2011). Productivity and priming: Morphemic decomposition in Arabic. *Language and Cognitive Processes, 26*(4–6), 624–652. https://doi.org/10.1080/01690965.2010.521022

Boudelaa, S., & Marslen-Wilson, W. D. (2015). Structure, form, and meaning in the mental lexicon: Evidence from Arabic. *Language, Cognition and Neuroscience, 30*(8), 955–992. https://doi.org/10.1080/23273798.2015.1048258

Boudelaa, S., Pulvermüller, F., Hauk, O., Shtyrov, Y., & Marslen-Wilson, W. (2010). Arabic morphology in the neural language system. *Journal of Cognitive Neuroscience, 22*(5), 998–1010. https://doi.org/10.1162/jocn.2009.21273

Boudelaa, S., Norris, D., Mahfoudhi, A., & Kinoshita, S. (2019). Transposed letter priming effects and allographic variation in Arabic: Insights from lexical decision and the same-different task. *Journal of Experimental Psychology Human Perception and Performance, 45*, 729–757. https://doi.org/10.1037/xhp0000621

Boudelaa, S., Perea, M., & Carreiras, M. (2024). Are the early stages of orthographic processing universal? Insights from masked priming with Semitic words. *Psychonomic Bulletin & Review*. Advance online publication. https://doi.org/10.3758/s13423-024-02563-8

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments & Computers, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental psychology, 58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

Brysbaert, M., Martínez, G., & Reviriego, P. (2024). *Moving beyond word frequency based on tally counting: AIgenerated familiarity estimates of words and phrases are a better index of language knowledge.* ResearchGate. Retrieved from https://www.researchgate.net/publication/383466866

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8*(3), 531–544. https://doi.org/10.3758/BF03196189

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS one, 5*(6), e10729. https://doi.org/10.1371/journal.pone.0010729

Carreiras, M., & Perea, M. (2004). Naming pseudowords in Spanish: Effects of syllable frequency. *Brain and Language, 90*(1–3), 393–400. https://doi.org/10.1016/j.bandl.2003.12.003

Carreiras, M., Alvarez, C. J., & de Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language, 32*(6), 766–780. https://doi.org/10.1006/jmla.1993.1038

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 857–871.

Chang, A., Zhu, X., Monga, A., Ahn, S., Srinivasan, T., & Thomason, J. (2023). Multimodal speech recognition for language-guided embodied agents. *Proceedings of Interspeech, 2023*, 1608–1612. https://doi.org/10.21437/Interspeech.2023-2262

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica, 32*(2), 133–143.

Darwish, K., Abdelali, A., Mubarak, H., & Eldesouki, M. (2020). *Arabic diacritic recovery using a feature-rich BiLSTM model.* arXiv preprint arXiv:2002.01207. https://doi.org/10.48550/arXiv.2002.01207

Dimitropoulou, M., Duñabeitia, J., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behaviour: The case of Greek. *Frontiers in Psychology, 1*(218), 1–12.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45*, 1246–1258. https://doi.org/10.3758/s13428-013-0326-1

Faizal, S. S., Khattab, G., & McKean, C. (2015). The Qur'an Lexicon Project: A database of lexical statistics and phonotactic probabilities for 19,286 contextually and phonetically transcribed types in Qur'anic Arabic. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, The University of Glasgow

Faust, M., & Lavidor, M. (2003). Semantically convergent and semantically divergent priming in the cerebral hemispheres: Lexical decision and semantic judgment. *Cognitive Brain Research, 17*(3), 585–597. https://doi.org/10.1016/S0926-6410(03)00172-1

Ford, M. A., Davis, M. H., & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language, 63*(1), 117–130. https://doi.org/10.1016/j.jml.2009.01.003

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning & Verbal Behavior, 12*(6), 627–635. https://doi.org/10.1016/S0022-5371(73)80042-8

Frost, R., Forster, K. I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked-priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 829–856. https://doi.org/10.1037/0278-7393.23.4.829

Frost, R., Deutsch, A., & Forster, K. I. (2000a). Decomposing morphologically complex words in a nonlinear morphology. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 751–765. https://doi.org/10.1037/0278-7393.26.3.751

Frost, R., Deutsch, A., Gilboa, O., Tannenbaum, M., & Marslen-Wilson, W. (2000b). Morphological priming: Dissociation of phonological, semantic, and morphological factors. *Memory & Cognition, 28*(8), 1277–1288. https://doi.org/10.3758/bf03211828

Frost, R., Kugler, T., Deutsch, A., & Forster, K. I. (2005). Orthographic structure versus morphological structure: Principles of lexical organization in a given language. *Journal of Experimental Psychology Learning, Memory, and Cognition, 31*(6), 1293–1326. https://doi.org/10.1037/0278-7393.31.6.1293

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*(3), 336–360. https://doi.org/10.1016/0749-596X(90)90004-J

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods, 48*, 963–972.

Gomez, P., & Silin, S. (2012). Visual word recognition models should also be constrained by knowledge about the visual system. *Behavioral and Brain Sciences, 35*(5), 25. https://doi.org/10.1017/s0140525x12000179

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology General, 136*(3), 389–413. https://doi.org/10.1037/0096-3445.136.3.389

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*(2), 228–244. https://doi.org/10.1016/0749-596X(90)90074-A

Grainger, J., & Hannagan, T. (2012). Explaining word recognition, reading, the universe, and beyond: A modest proposal. *Behavioral and Brain Sciences, 35*(5), 288–289. https://doi.org/10.1017/S0140525X12000064

Grainger, J., & Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology, 2*(54), 1–13. https://doi.org/10.3389/fpsyg.2011.00054

Heath, J. (2003). Arabic derivational ablaut, processing strategies, and consonantal "roots". In J. Shimron (Ed.), *Language processing and acquisition in languages of Semitic root-based, morphology* (pp. 100–115). Amsterdam: John Benjamins.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*, 643–650. https://doi.org/10.3758/BRM.42.3.643

Khoja, S., & Garside, R. (1999). *Stemming Arabic Text*. Lancaster University.

Kučera, M., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Kuperman, V., Rastle, K., & Davis, M. H. (2022). Morphological processing and its impact on reading comprehension: A cross-linguistic perspective. *Psychological Science, 33*(5), 842–857.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lane, H. B., Gutlohn, L., & van Dijk, W. (2019). Morpheme Frequency in Academic Words: Identifying High-Utility Morphemes for Instruction. *Literacy Research and Instruction, 58*(3), 184–209. https://doi.org/10.1080/19388071.2019.1617375

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–282). ACM. https://doi.org/10.1145/564376.564428

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101*(1), 3–33. https://doi.org/10.1037/0033-295X.101.1.3

McNeil, K. & Miled, F. (2010). Tunisian Arabic Corpus (TAC): 895,000 words. Retrieved from http://www.tunisiya.org

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. K., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science, 320*, 1191–1195.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 148–197). Lawrence Erlbaum Associates Inc.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General, 118*(1), 43–71. https://doi.org/10.1037/0096-3445.118.1.43

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews, 82*, 591–605.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. *LREC, 14*, 1094–1101.

Perea, M., & Carreiras, M. (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 134–144. https://doi.org/10.1037/0096-1523.24.1.134

Perea, M., Abu Mallouh, R., & Carreiras, M. (2010). The search of an input coding scheme: Transposed-letter priming in Arabic. *Psychonomic Bulletin and Review, 17*, 375–380.

Perea, M., Abu Mallouh, R., García-Orza, J., & Carreiras, M. (2011). Masked priming effects are modulated by expertise in the script. *Quarterly Journal of Experimental Psychology, 64*(5), 902–919. https://doi.org/10.1080/17470218.2010.512088

Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word-identification times in young

readers. *Journal of Experimental Child Psychology, 116*, 37–44. https://doi.org/10.1016/j.jecp.2012.10.014

Perea, M., Abu Mallouh, R., & Carreiras, M. (2014). Are root letters compulsory for lexical access in Semitic languages? The case of masked form-priming in Arabic. *Cognition, 132*(3), 491–500. https://doi.org/10.1016/j.cognition.2014.05.008

Prunet, J. F., Beland, R., & Idrissi, A. (2000). The mental representation of semitic word. *Linguistic Inquiry, 31*, 609–648. https://doi.org/10.1162/002438900554497

Robson, J., Pring, T., Marshall, J., & Chiat, S. (2003). Phoneme frequency effects in jargon aphasia: A phonological investigation of nonword errors. *Brain and Language, 85*(1), 109–124. https://doi.org/10.1016/s0093-934x(02)00503-5

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H. D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Kwon, N., Lõo, K., Marelli, M., …, & Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, *54*(6), 2843–2863. https://doi.org/10.3758/s13428-021-01772-6

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., Comesaña, M., & Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology, 68*(4), 680–696. https://doi.org/10.1080/17470218.2014.964271

Stenneken, P., Conrad, M., Hutzler, F., Braun, M., & Jacobs, A. M. (2005). Frequency effects with visual words and syllables in a dyslexic reader. *Behavioural Neurology, 16*(2–3), 103–117. https://doi.org/10.1155/2005/427605

Taft, M., & Zhu, X. (1997). Submorphemic processing in reading Chinese. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(3), 761–775. https://doi.org/10.1037/0278-7393.23.3.761

Thorndike, E. L. (1921). *The teacher's word book*. Teachers College, Columbia University.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Columbia University.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014a). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014b). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of Experimental Psychology, 67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Yonelinas, A. P. (2002). Components of episodic memory: The contribution of recollection and familiarity. In A. Baddeley, J. P. Aggleton, & M. A. Conway (Eds.), *Episodic memory: New directions in research* (pp. 31–52). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198508809.003.0003