BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Transferring Buckwalter Transcription to a Batch Mode: A Method to Making It More Accessible

Ibrahim Abdulrahman Alluhaybi , ***Talal Musaed Alghizzi*** *

Department of English Language and Literature, College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

ABSTRACT

This study addresses the challenges associated with the manual application of the Buckwalter Arabic Transcription System, a pivotal tool in computational linguistics for representing Arabic script using ASCII characters. Although the system ensures high fidelity and reversibility, its usability is hindered by its non-phonetic nature, steep learning curve, and reliance on manual referencing. To enhance accessibility and usability, this research introduces a web-based batch-mode interface that automates the Buckwalter transcription process. The tool allows users to input Arabic text and instantly receive standardized Buckwalter transliterations alongside International Phonetic Alphabet (IPA) representations. This dual-output approach supports a wide range of applications in linguistics, education, and natural language processing. The study explores the theoretical and linguistic foundations of the Buckwalter system, outlines its strengths and weaknesses, and analyzes its morphological implications. It further presents practical examples using real-world data, including excerpts from the Universal Declaration of Human Rights. The batch-mode website (ipabwat.com) streamlines the transcription of large Arabic texts, offering downloadable results in CSV format and an intuitive interface suited for both novice and expert users. By integrating automation with linguistic precision, the tool eliminates the need for manual chart referencing and reduces transcription errors, thus broadening the scope of Arabic text processing. Ultimately, this work aims to democratize access to Arabic computational tools, making the Buckwalter system more functional for researchers, developers, and

*CORRESPONDING AUTHOR:

Talal Musaed Alghizzi, Department of English Language and Literature, College of Languages and Translation, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia; Email: Tmalghizzi@imamu.edu.sa

ARTICLE INFO

Received: 5 May 2025 | Revised: 11 June 2025 | Accepted: 24 June 2025 | Published Online: 21 July 2025
DOI: https://doi.org/10.30564/fls.v7i7.9440

CITATION

Alluhaybi, I.A., Alghizzi, T.M., 2025. Transferring Buckwalter Transcription to a Batch Mode: A Method to Making It More Accessible. Forum for Linguistic Studies. 7(7): 992–1004. DOI: https://doi.org/10.30564/fls.v7i7.9874

learners across disciplines. It represents a critical step forward in enhancing the usability and reach of Arabic linguistic technologies.

Keywords: Buckwalter Transcription; Arabic NLP; Batch Transliteration; ASCII Encoding; Computational Morphology

1. Introduction

In the effective computational processing of natural languages, the Arabic language represents a significant challenge to designers and programmers due to its unique morphology, phonetics, phonology, and non-standardized orthography[1]. Various attempts and efforts to develop Arabic morphological stemmers and analyzers have been proposed, but were characterized by their reliance on hand-crafted rules and limited coverage[2]. However, one of the most successful solutions was that of Buckwalter, who proposed the Buckwalter Transcription chart/analyzer[3, 4]. According to Buckwalter, the transcription was developed to provide a standardized machine-readable representation of Arabic text that could be used in various language processing tasks[3], and to facilitate Arabic morphological analysis, disambiguation, machine translation[4], information retrieval, and text classification[5]. It is true that such transcription is deemed beneficial to its designated users, but there are some challenges that these users need to overcome in order to apply the transcription effectively. For instance, Habash et al. observed that Buckwalter's ASCII transliteration system may introduce a degree of ambiguity due to its limited symbol set and lack of phonological transparency, which can make it somewhat challenging for users when selecting the correct code mapping[6]. In particular, users may struggle to memorize the various codes used to represent Arabic characters and diacritics, which often leads them to rely on reference charts and other resources for accurate execution of the transcription. The researchers also emphasized that the Buckwalter system does not always provide a one-to-one mapping between Arabic characters and their corresponding codes, and this complicates the transcription process. For instance, some diacritics may be represented by multiple codes depending on their position within a word or a sentence, which is likely to create some hurdles for users in specifing which code to use without immediate access to a reference chart. Finally, regardless of the transcriptions' effectiveness, it still requires significant effort and familiarity for effective use, especially for users who are not already well-versed in the language and its conventions. Therefore, we undertook the task of overcoming these issues by transferring Buckwalter Transcription to an online batch mode, which will likely make it easier, more accessible, and more convenient to apply for users of different language backgrounds.

2. Literature Review

To set the foundation for our proposed solution, this section provides an overview of the broader computational and linguistic background of Arabic language processing. This includes an understanding of the unique structural characteristics of the Arabic language, the limitations of existing morphological analysis tools, and the impact of Buckwalter Transcription on Arabic linguistics. To do that, we review the foundational work on the characteristics of the Arabic language, the issues in processing different forms, as well as prior work on enhancing transcription tools, in order to lay the groundwork for the development of our batch-mode web interface.

2.1. Arabic Language Characteristics and Processing Technology

The Arabic language is a linguistic continuum[7], Arabic exhibits a dual linguistic landscape, with one prominent pole representing the standardized Arabic used in written and formal spoken contexts. In contrast, a cluster of interconnected Arabic dialects is characterized by significant phonological, morphological, syntactic, and lexical distinctions among themselves and in comparison to the standardized written forms. Maamouri et al. assert that this poses challenges for developing speech-to-text systems in Arabic, as the spoken dialects lack official written standards despite their ongoing expansion[8]. A significant level of linguistic diversity exists, resulting in multiple forms that are difficult to detect and regroup. Recent research has further explored how

dialectal variance in Arabic hinders NLP models in classification and interpretation tasks[9]. The imperative need to foster effective communication with Arabic-speaking individuals underscores the significance of developing Natural Language Processing (NLP) systems tailored for this purpose, aligning applications with the nuances and characteristics of the Arabic language[10] in order to ensure optimal linguistic processing. Nevertheless, progress in Arabic NLP lags behind that of other languages like English and Chinese[11, 12]. Fadel et al. highlight that this disparity can be attributed to a number of factors, including limited investments and inadequate linguistic resources dedicated to Arabic NLP, which pose significant challenges for researchers and developers[11, 13]. The Buckwalter Transcription System, developed by Tim Buckwalter, plays a pivotal role in computational linguistics and Arabic language processing[14]. It offers a method to encode Arabic text into a format that is computationally accessible, facilitating natural language processing tasks such as information retrieval and text analysis. Moreover, we present a cutting-edge website that streamlines the Buckwalter transcribing process. This distinctive platform utilizes the latest advancements in Natural Language Processing (NLP) to offer linguists and researchers an intuitive, user-friendly tool for Arabic text transcription, delivering exceptional ease and accuracy.

2.2. Types of Language Transcriptions

According to Gorgis, transcription pertains to a written representation of spoken language, capturing a language's phonological or morpho-phonological aspects[15]. It translates spoken language into written form, as described by Beesley. Phonetic transcription and orthographic transcription are the two main types of language transcription. Abdul-Mageed et al. emphasize that spoken language's phonetic and phonological properties are the main focus of phonetic transcription[16]. Using established character sets such as the International Phonetic Alphabet, phonetic transcription uses specific criteria to translate individual speech sounds or phones into written symbols.

In contrast, Habash et al. maintain that orthographic transcription records how words are spelled according to the grammar rules and conventions of the language's writing system, this type of transcription focuses on the visual representation of language[6]. While orthographic transcription represents spoken words in written form according to a language's orthographic conventions, phonetic transcription concentrates on speech sounds and employs specific character sets like the International Phonetic Alphabet.

Transliteration refers to a method of writing where specific orthographic symbols are meticulously substituted by establishing a direct and entirely reversible correspondence with the customary orthography of the language, as Gorgis states[15]. This form of transliteration, often termed 'strict transliteration' or 'orthographic transliteration', is essential to linguistic studies[6]. The Buckwalter transliteration system is an ASCII-based method that faithfully represents Arabic orthography in one-to-one correspondence[3, 4]. Nawar points out that, in contrast to typical Romanization systems, which often introduce additional morphological details not presents in the Arabic script, the Buckwalter Arabic transliteration adheres to the standard encoding conventions used for computer representation of Arabic characters[17]. Habash et al. explain that Arabic characters can take up to four distinct forms, with the specific shape of each character determined by its position within a word; the positional types are initial, medial, final, and isolated[18]. Arabic script incorporates all five long vowels alongside the consonants within standard writing. However, the three short vowels and the "sukun", which indicats the absence of a vowel sound between consonants, are not visibly represented in the written form.

Moreover, Habash et al. add that there are 28 letters and eight diacritical symbols in the standard Arabic alphabet[6]. Eight extra symbols can be used as distinct letters or in unique combinations with other diacritics and letters. An illustrative instance is the Hamza, which can function both as an independent letter (ء) and as a diacritics in combination with other letters like ؤ , أ, and ئ . Consequently, it is feasible to establish an orthographic symbol system for Arabic in which the Hamza is considered not solely a letter but also a diacritic with a restricted set of combinations. Conventional computer encodings for Arabic, like CP1256, ISO-8859, and Unicode2, do not adopt this approach. They consider the additional eight symbols as separate letters[19]. Transliteration aims to portray the characters, ideally correctly and clearly, rather than the sounds of the source language. Contrarily, transcription aims to record sounds rather than writing. In their paper, Habash and Roth noted that prior attempts to create a comprehensive and easily legi-

ble one-to-one transliteration system for the Arabic script, while maintaining compatibility with Arabic computer encodings, have been notably scarce [20]. Many previously devised schemes for representing Arabic characters to Western readers have primarily focused on conveying phonological and morphological aspects, often leaning toward transcription. Some even straddle the line between phonology and orthography, occasionally making exceptions for the transliteration of specific morphemes, such as the definite article. Hajic argues that one notable example, ISO 233—an international standard developed by the International Organization for Standardization (ISO) for the Romanization of Arabic characters—comes close to achieving our desired goal but needs greater consistency with computer encodings [21]. ISO 233 provides a standardized method for transliterating Arabic script into Latin characters, ensuring consistency in the representation of Arabic names, places, and terms when using the Latin alphabet. This standard is crucial in contexts where accuracy and consistency in representing Arabic names or terms in documents, databases, or international communications are essential.

2.3. Origin, Rationale, and Creation of Buckwalter Transcription

Broselow et al. contended that, under the guidance of Ken Beesley, the Buckwalter Arabic Transliteration System was first developed in 1988 as a component of the ALP-NET Arabic Project [22]. Subsequently, the Buckwalter Transcription System, conceived by Tim Buckwalter and Derek Foxley, was developed during the late 1990s. The the Buckwalter Transcription System was created provide a consistent, machine-readable format for Arabic text, suitable for a wide range of tasks in natural language processing [3]. This system is central to computational linguistics, offering a standardized tool for scholars and professionals to work with Arabic language data. Its development has significantly streamlined the examination and handling of Arabic text, making notable contributions to various areas that depend on accurate and well-organized linguistic data.

Arabic presents substantial barriers in natural language processing. This is primarily due to its complex morphology and lack of standardized spelling [1]. The language's intricate morphology necessitates use of sophisticated algorithms for proper analysis. Furthermore, the need for standardized spelling complicates computing systems even further. Previous efforts to construct Arabic morphological analyzers and stemmers relied on manually generated rules and suffered from limitedcoverage [2]. These methods need to capture the full range of Arabic linguistic complexities. Overcoming these obstacles needs novel methodologies and resources, emphasizing the crucial need for Arabic computational linguistics research [1].

The creation of the Buckwalter Transcription System was a watershed moment in Arabic computational linguistics. Tim Buckwalter devised this standardized transliteration approach in 2002 to transliterate the Arabic text in a manner that machines could handle effectively [3]. The system was rigorously designed with Arabic morphology in mind to enable exact morphological analysis, disambiguation, and machine–translation tasks [4]. The Buckwalter system reduced computational operations by encoding Arabic text in a machine-readable format, allowing more accurate and efficient linguistic analysis. Its design was tailored to the linguistic peculiarities of Arabic, overcoming the difficulties posed by the language's complicated morphology and nonstandardized orthography.

Consequently, the system emerged as an indispensable resource for scholars and professionals involved in various activities, from scrutinizing word structures to automating language translation [3]. This transformation significantly reshaped the landscape of Arabic language processing, solidifying the system's position as a fundamental component in the discipline. It underscores the crucial significance of uniform transliteration techniques in propelling Arabic computational linguistics to new heights.

The Buckwalter Transcription System was a critical component of a more significant attempt to improve Arabic language processing. This included building robust morphological analyzers, part-of-speech taggers, and parsers [23]. These complementary technologies constitute the backbone of computational linguistics, allowing the thorough analysis and comprehension of Arabic text. The Buckwalter Transcription system enabled seamless integration with these advanced language technologies as a standardized transliteration approach. Recent efforts have also focused on developing large annotated corpora tailored for dialectal variation, which further reinforce the need for consistent and reversible transcription systems such as Buckwalter Transcription Sys-

tem [24]. This collaborative work has been crucial in enhancing Arabic natural language processing capabilities, leading to a better understanding of Arabic linguistics and boosting various computational applications.

The Buckwalter Arabic Transliteration System is widely used due to its critical role in various natural language processing tasks involving Arabic text. Its applications range from machine translation to information retrieval and text categorization [5, 25]. Although the Buckwalter Transcription System is designed for standard Arabic, it can technically represent any text written in Arabic script. This orthographic representation allows the system to transcribe any Arabic-script text, regardless of whether it reflects Standard Arabic or dialectal variations. The system's standardized representation of Arabic text in machine translation promotes uniformity in transliteration, a critical step in adequately transforming text from Arabic to other languages [26]. This consistency reduces errors and improves the overall quality of the translation results. Furthermore, in the field of information retrieval, where search engines' success depends on correct text representation, the Buckwalter method provides a dependable way of indexing and retrieving Arabic materials. Its use ensures that queries and documents are correctly matched, resulting in more relevant results.

The system was a critical component of a more significant attempt to improve Arabic natural language processing [27]. As emphasized by Habash et al., Tim Buckwalter's approach was crucial for enabling diacritization, labeling, and parsing of Arabic text [1]. As observed by Habash and Rambow, it played a significant role in streamlining essential linguistic processes such as tokenization, part-of-speech tagging, and morphological disambiguation, thereby combining these tasks into a continuous method. Furthermore, the Buckwalter Transcription System emerged as a pivotal contribution to the production of annotated Arabic corpora, a critical resource for developing and evaluating natural language processing systems [2]. The Buckwalter Transcription System's diverse impact on Arabic language processing emphasizes its significance in defining the procedures and instruments required for accurate and successful linguistic analysis. The Buckwalter Transcription System is a cornerstone in the evolution of Arabic natural language processing because of its contributions to diacritization, tagging, parsing, and corpus building.

2.4. Strengths of Buckwalter Transcription

The Buckwalter Arabic Transliteration System's utilization of ASCII characters significantly enhanced the ease of processing Arabic text in computational systems, particularly during the nascent stages of computing when support for non-Latin scripts was constrained [1, 28]. Systems were first built to handle the English alphabet and a limited set of ASCII characters. This posed a significant difficulty for languages with unique scripts, such as Arabic. Alternatively, the Buckwalter method efficiently overcame this barrier by transliterating Arabic text into ASCII letters. ASCII characters are recognized and processed by all computer systems and software programs. At a period when direct support for non–Latin scripts was limited, this pragmatic adaptation enabled the smooth integration of Arabic text into computational activities such as information retrieval and natural language processing. Furthermore, because Arabic text was compatible with ASCII letters, it could be easily integrated into existing computing systems without requiring significant adjustments or specialized software.

The Buckwalter Arabic Transliteration System offers a significant benefit in that it can be converted back to the original Arabic script without losing any information [3]. This property is extremely important in both linguistic and computational contexts. When dealing with transliterated text, it is sometimes necessary to return to the original script for thorough examination or to provide results in a format that follows standard linguistic conventions. This reversibility is essential for morphological analysis, syntax, and phonology tasks. Researchers and linguists use this feature to undertake comprehensive linguistic investigations. In applications such as machine–translation or natural language processing, where precision in processing and creating Arabic text is critical, reverting to the original script guarantees that the language's nuances and subtleties are retained, as explained by Habash [29]. Furthermore, this reversibility contributes to the integrity of textual data. When transliterated text must be shared, preserved, or processed by several systems, the Buckwalter Transliteration System's ability to return to the original Arabic script ensures that various stakeholders keep the content intact and interpretable.

The Buckwalter Arabic Transliteration System distinguishes itself from other transcription systems by including

representations for Arabic diacritics and short vowels. This inclusion is critical in preserving the subtle intricacies of Arabic script's[8]. Diacritics and short vowels have great linguistic importance in Arabic orthography, offering important phonological and grammatical information. They distinguish between homographs, aid pronunciation, and provide contextual cues for appropriate comprehension. This trait is critical in linguistic research and computational activities, notably morphological analysis and part-of-speech tagging. It enables more accurate text analysis, resulting in more exact findings. Furthermore, diacritics are valuable aids in Arabic language learning in educational settings, helping students accurately pronounce and understand meanings. The conservation of diacritics provides a faithful representation of the original Arabic text in many computational applications, such as machine translation, sentiment analysis, and information retrieval, where context and accuracy are critical. This, in turn, leads to more accurate and contextually relevant results.

The Buckwalter Transliteration System received practical application in experimental psycholinguistics, particularly in visual word recognition via DMDX. The ASCII encoding of Buckwalter Transliteration System provides a precise and consistent input format for Arabic stimuli. For example, Alluhaybi and Witzel used Buckwalter transliteration to encode Arabic stimuli for transcription purposes in visual word recognition. They utilized DMDX to perform lexical decision tasks to observe measurable effects such as word chunks and letter connectedness on visual word recognition[30]. In addition, Witzel et al. used Buckwalter transliteration to transcribe Arabic primes and targets before presenting them in DMDX, reporting comparable response times and error rates, which emphasizes the encoding's reliability in tightly controlled visual word recognition paradigms[31].

2.5. Weaknesses of the Buckwalter Transcription System

The Buckwalter Arabic Transliteration System, while extremely useful for computational processing, has one notable feature: it is not phonetic. This means that it may fail to correctly match Arabic word pronunciation, thus posing a barrier for those unfamiliar with the complexities of the Arabic language[3]. Unlike phonetic transcription methods, which strive to reflect genuine speech sounds, the Buckwalter approach concentrates on orthographic representation, adhering closely to the written form of the Arabic script. As a result, this technique may not provide a natural bridge for users who rely on phonetic signals to acquire or understand a language. This non-phonetic characteristic may lead to mispronunciations or misunderstandings, especially for students or researchers who rely on accurate phonetic representations. However, it is vital to emphasize that the system's primary goal is to preserve the visual structure of Arabic text in computational applications. The Buckwalter Transcription System succeeds in this regard by preserving a one-to-one correlation between Arabic script and its transliteration, guaranteeing correctness and consistency for tasks such as morphological analysis and information retrieval.

The Buckwalter Arabic Transliteration System, while necessary for Arabic language processing, has a significant limitation: it is designed solely for Arabic, making it less adaptable to other Semitic languages with comparable scripts[3]. This limitation stems from its strict adherence to Arabic orthographic standards and writing conventions. While Arabic and other Semitic languages, such as Hebrew and Amharic, share script roots, each exhibits distinct phonological and orthographic characteristics. As a result, applying the Buckwalter approach to these languages may result in errors and misrepresentations. This concern has also been echoed in recent dialect detection studies, where morphological ambiguity leads to misclassification[32]. Specific phonetic characteristics present in Arabic, for example, may not exist in other Semitic languages, thus leading to transliteration problems. Consequently, the Buckwalter system's value for scholars and researchers working with numerous Semitic languages may be limited. It is critical, however, to recognize that the system's specialized focus on Arabic enables it to excel at tasks specific to this language, such as morphological analysis and part-of-speech tagging.

Individuals unfamiliar with the Arabic alphabet may find using the Buckwalter Arabic Transliteration System challenging, perhaps resulting in a steep learning curve[3]. Because the system preserves the visual qualities of Arabic, people unfamiliar with the script may initially be perplexed. The complexities of Arabic orthography, such as its particular letter forms and script orientation, may take considerable time to master. Furthermore, although diacrit-

ics and short vowels are suitable for linguistic correctness, they may present additional challenges for novices. However, it is crucial to emphasize that uers may overcome this initial barrier with dedicated work and appropriate resources. Learning tools, tutorials, and practice activities are readily available to help users gain proficiency with the Buckwalter system. Once mastered, the system is a significant tool for various computational linguistic tasks related to Arabic text processing.

The Buckwalter Transcription System is critical for overcoming the inherent challenges presented by the morphological complexity of the Arabic language. The removal of short vowels and orthographic diacritics in Arabic—noted for its extensive system of affixation and clitics—sometimes leads to uncertainty in morphological analysis [3]. An Arabic Treebank (ATB) word has, on average, around two potential morphological analyses, demonstrating its complexity. Consider the word "++w", which may be analyzed as either a noun with no affixes or as a conjunction prefix and a pronominal possessive suffix, "+ +y'my."

Tokenization and morphological tagging—including part-of-speech (POS) tagging—are combined operations in the Buckwalter Transcription System. Using a morphological analyzer, the system first generates a list of all feasible analyses each word in a given phrase. This is followed by applying classifiers to the words for ten morphological features, including clitics and affixes. A thorough procedure is used to train and decode these classifiers.

Finally, using the output of the classifiers, the system picks the most suitable analysis from the possibilities supplied by the morphological analyzer. Classifiers may only partially disambiguate the possibilities or offer contradictory information, making this a complicated process. This approach yields the original text, with each word enhanced by values for numerous morphological features, resulting in thorough morphological disambiguation. Furthermore, these properties provide information on the existence of clitics and affixes, enabling successful tokenization. Moreover, based on the morphological information, the system derives the POS tag, accomplishing tokenization, standard POS tagging, and complete morphological disambiguation. The Buckwalter Transcription System handles the rich morphological properties Arabic text, allowing precise linguistic analysis and computational processing.

Guidelines for mapping Arabic sounds with precision, according to Habash et al. [6]:

1. Mapping letters to sounds in Arabic is generally straightforward, especially for most consonants. Some consonants are familiar to English speakers, resulting in identical transcription and transliteration. However, the Arabic consonant Hamza presents various forms based on complex spelling rules that depend on the surrounding vowels. Nevertheless, these different forms yield the same pronunciation.
2. Arabic employs three short vowel diacritics, symbolized by the letters a, u, and i. Additionally, three nunation diacritics represent short vowels followed by an /n/ sound. It is important to note that these nunation diacritics do not indicate nasalized vowels. In Arabic, long vowels and diphthongs are indicated by combinations of a short vowel and a consonant, as Darwish stated [33].
3. The letter Alif (A) indicates the long vowel /ā/ at a word's outset, while it also marks a few morphophonemic symbols where the Alif itself is not pronounced.
4. The /tā' marbūTa/ ending is typically associated with the feminine form in Arabic. It is exclusively permitted to follow a word and requires a diacritic [6]. In standard Arabic pronunciation, it is generally articulated as /t/ unless it appears after a diacritic, in which case it remains silent.
5. The /alif maqṣūra/ symbol, represented by ى or ý, signifies a dotless ي (y). In standard Arabic, it remains silent and consistently appears after a short vowel and at the end of a word.

For making the Buckwalter transliteration table, these considerations were made:

1. In Arabic, each letter or sound corresponds to a single English character. However, in some cases, specific Arabic characters produce a sound equivalent to two English letters, necessitating their representation by a single letter or a shared symbol.
2. The table must be entirely mnemonic, meaning that each element within it is linked to either (a) the sound of the Arabic letter [34], (b) a physical characteristic of the original Arabic letter, or (c) the name by which it

is recognized, as illustrated by Table 1 below (adapted from the Wikimedia Foundation [35]).

Table 1. Mnemonic Mapping of Arabic Letters to Buckwalter Transliteration Symbols with Corresponding IPA Pronunciations.

Arabic Letters	Buckwalter	Pronunciation
ا	A	[aː]
ب	B	[b].
ت	T	[t].
ث	V	[t].
ج	J	[dʒ]
ح	H	[ħ]
خ	X	[x].
د	D	[d].
ذ	*	[ð]
ر	R	[r]
ز	Z	[z]
س	S	[s].
ش	$	[ʃ].
ص	S	[sˤ]
ض	D	[dˤ]
ط	T	[tˤ]
ظ	Z	[ðˤ∼zˤ]
ع	E	[ʕ]
غ	G	[ɣ∼ʁ]
ف	F	[f]
ق	Q	[q]
ك	K	[k]
ل	L	[l]
م	M	[m]
ن	N	[n]
ه	H	[h]
و	W	[w]
ي	Y	[aː]

3. Buckwalter Transcription Batch-Mode Website

To demonstrate how the theoretical framework and linguistic insights underlying Buckwalter transcription can be leveraged in practice, we implemented a dedicated batch-mode web interface (ipabwat.com). Building on our analysis of standardized character mapping and consistency in transliteration, the site translates these mechanisms into three core features:

1. Bulk-paste input of Modern Standard Arabic text (supports very large passages).
2. Simultaneous display of both ASCII Buckwalter transliteration and IPA transcription after conversion.
3. One-click processing that renders results in an on-page table, with an option to download the output as a CSV file.

Behind the scenes, the engine applies the one-to-one mappings from "The Buckwalter Transliteration System" while the interface layout supports each core conversion step described above.

The website is designed to provide a user-friendly platform for converting Arabic script into the standardized Buckwalter Arabic Transliteration. Users can easily receive the matching representation in Latin characters by inputting Arabic text. The correctness and uniformity of this conversion procedure ensure that the subtleties of the original Arabic text are properly retained in the resultant transliteration. This website provides a dependable and effective solution for linguistic study, natural language processing, and any other application that requires a Latinized representation of Arabic text. This useful application is easily accessible to anyone who wants to improve their work with Arabic language data in a number of scenarios.

3.1. Interface

The website features a straightforward and user-friendly interface design, with the text input box at the top and the conversion button beneath it (see **Figure 1**). The language selection radio button is located beneath the conversion button and now offers only one language option, Arabic.

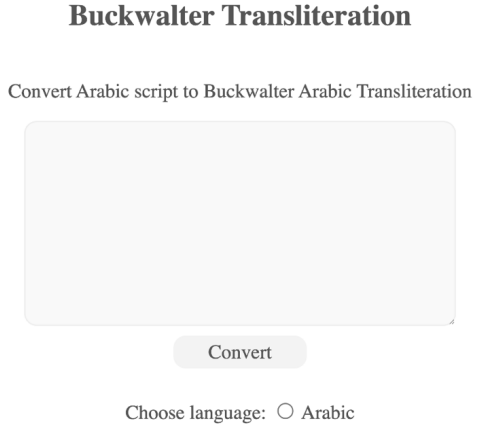

Figure 1. Web-Based Interface for Buckwalter Transliteration.

The first step is to enter an Arabic script in the text field stated (see **Figure 2**). For example, مساء الخير. This prompts the user to choose the language, as highlighted in yellow below.

Figure 2. User Input of Arabic Script and Language Selection.

After choosing the language of choice—Arabic in this case—a tools subheading will appear. The tool is "Buckwalter/IPA",as shown below in **Figure 3**.

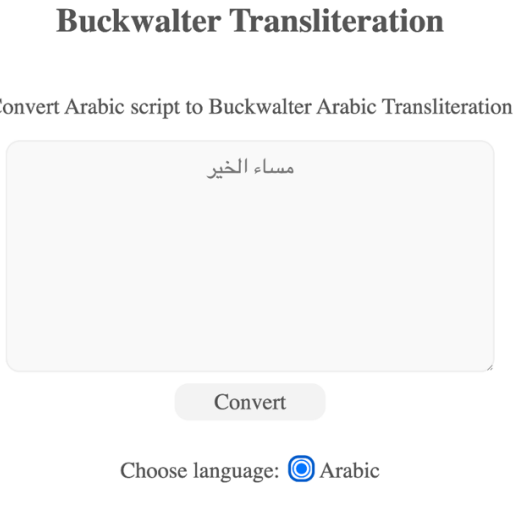

Figure 3. Tool Selection for Buckwalter/IPA Transliteration.

When you choose the "Buckwalter/IPA" option from the tools menu and start the conversion, a table with three separate columns appears. The first column, labelled "Arabic Script," reveals the user's native script. The second column, labelled "Buckwalter Transliteration," displays the text produced in the standardized Buckwalter format. Finally, the third column, labelled "IPA", the International Phonetic Alphabet method is used to precisely reflect word pronunciation. This detailed table gives readers a clear and structured overview of the Arabic text, its Buckwalter transliteration, and the accurate phonetic representation of the words/phrases/texts. The site also offers a option to download the file in the form of a sheet(.csv). Below in **Figure 4** is an image of the generated text from the phrase above.

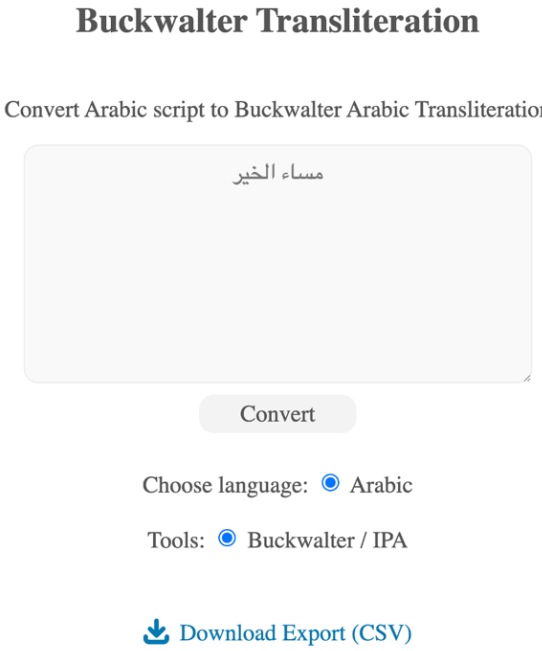

Figure 4. Output Table Showing Buckwalter Transliteration and IPA Representation.

3.2. Example of a Long Arabic Text

For a long Arabic text example, we chose article number 2 posted in United Nations' website. (2023, available from: https://www.un.org/ar/about-us/universal-declaration-of-human-rights/). See **Figure 5**.

Arabic Script	Buckwalter transliteration	IPA
لكلِّ إنسان حقُّ التمتُّع بجميع الحقوق والحرِّيات المذكورة في هذا الإعلان، دونما تمييز من أيِّ نوع، ولا سيما التمييز بسبب العنصر، أو اللون، أو الجنس، أو اللغة، أو الدِّين، أو الرأي سياسيًّا وغير سياسي، أو الأصل الوطني أو الاجتماعي، أو الثروة، أو المولد، أو أيِّ وضع آخر. وفضلاً عن ذلك لا يجوز التمييزُ علي أساس الوضع السياسي أو القانوني أو الدولي للبلد أو الإقليم الذي ينتمي إليه الشخص، سواء أكان مستقلاً أو موضوعًا تحت الوصاية أو غير متمتِّع بالحكم الذاتي أم خاضعًا لأيِّ قيد آخر على سيادته.	lkl~i <nsAn Hq~u Altmt~uE bjmyE AlHqwq wAlHr~iyAt Alm*kwrp fy h*A Al<ElAn، dwnmA tmyyz mn >y~i nwE، wlA symA Altmyyz bsbb AlEnSr، >w Allwn، >w Aljns، >w Allgp، >w Ald~iyn، >w Alr>y syAsyF~A wgyr syAsy، >w Al>Sl AlwTny >w AlAjtmAEy، >w Alvrwp، >w Almwld، >w >y~i wDE lxr. wfDlAF En *lk lA yjwz Altmyyzu Ely >sAs AlwDE AlsyAsy >w AlqAnwny >w Aldwly llbld >w Al<qlym Al*y yntmy <lyh Al$xS، swA' >kAn mstqlAF~ >w mwDwEFA tHt AlwSAyp >w gyr mtmt~iE bAlHkm Al*Aty >m xADEFA l>y~i qyd lxr ElY syAdth.	lkli ʔnsaːn ħqu ʔltmtuʕ bɣmiːʕ ʔlħquːq waːlħriiːaːt ʔlmðkuːrh fiː hðaː ʔlʔʕlaːn، duːnma: tmiːiːz mn ʔiːi nuːʕ، wlaː siːma: ʔltmiːiːz bsbb ʔlʕnsˤr، ʔuː ʔlluːn، ʔuː ʔlɣns، ʔuː ʔllyh، ʔuː ʔldiiːn، ʔuː ʔlrʔiː siːaːsiːaa: wyiːr siːaːsiː، ʔuː ʔlʔsˤl ʔluːtˤniː ʔuː ʔlaːɣtmaːʕiː، ʔuː ʔlθruːh، ʔuː ʔlmuːld، ʔuː ʔiːi wdˤʕ ʔxr. wfdˤlaːa ʕn ðlk laː jɣuːz ʔltmiːiːzu ʕliː ʔsaːs ʔluːdˤʕ ʔlsiːaːsiː ʔuː ʔlqaːnuːniː ʔuː ʔldu:liː llbld ʔuː ʔlʔqliːm ʔlðiː jntmiː ʔliːh ʔlʃxsˤ، suːaːʔ ʔkaːn mstqlaːa ʔuː muːdˤuːʕaa: tht ʔluːsˤaːiːh ʔuː yiːr mtmtiʕ baːlħkm ʔlðaːtiː ʔm xaːdˤʕaa: lʔiːi qiːd ʔxr ʕlaː siːaːdth.

Figure 5. Buckwalter and IPA Output for a Full Arabic Paragraph.

3.2.1. Arabic Version

مادة 2
لكلّ إنسان حقُّ التمتُّع بجميع الحقوق والحرّيات المذكورة في هذا الإعلان، دونما تمييز من أيّ نوع، ولا سيما التمييز بسبب العنصر، أو اللون، أو الجنس، أو اللغة، أو الدِّين، أو الرأي سياسيّ"ا وغير سياسي، أو الأصل الوطني أو الاجتماعي، أو الثروة، أو المولد، أو أيّ وضع آخر. وفضلاً عن ذلك لا يجوز التمييزُ علي أساس الوضع السياسي أو القانوني أو الدولي للبلد أو الإقليم الذي ينتمي إليه الشخص، سواء أكان مستقلا" أو موضوعًا تحت الوصاية أو غير متمتّع بالحكم الذاتي أم خاضعًا لأيّ قيد آخر على
سيادته.

3.2.2. The English Translated Version

The Arabic text is also translated into English on the United Nations website (available at Universal Declara-on of Human Rights | United Na-ons)

Article 2

Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional, or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing, or under any other limitation of sovereignty.

3.2.3. Buckwalter Transliteration

2. AlmAdp

*lkl i <nsAn Hq u Altmt uE bjmyE AlHqwq wAlHr iyAt Alm*kwrp fy h*A Al<ElAn ‘dwnmA tmyyz mn >y i nwE ‘wlA symA Altmyyz bsbb AlEnSr ‘>w Allwn ‘>w Aljns ‘>w Allgp ‘>w Ald iyn ‘>w Alr>y syAsyF A wgyr syAsy ‘>w Al>SlAlwTny >w AlAjtmAEy ‘>w Alvrwp ‘>w Almwld ‘>w >y i wDE |xr. wfDlAF En *lk lA yjwz Altmyyzu Ely >sAs AlwDE AlsyAsy >w AlqAnwny >w Aldwly llbld >w Al<qlym Al*y yntmy <lyh Al$xS ‘swA’>kAn mstqlAF >w mwDwEFA tHt AlwSAyp >w gyr mtmt iE bAlHkm Al*Aty >m xADEFA l>y i qyd |xr ElY syAdth.*

3.2.4. IPA

2. ʔlma:dh

lkli ʔnsa:n ħqu ʔltmtuʕ bʒmi:ʕ ʔlħqu:q wa:lħrii:a:t ʔlmŏku:rh fi: hða: ʔlʔʕla:n ‘du:nma: tmi:i:z mn ʔi:i nu:ʕ ‘wla: si:ma: ʔltmi:i:z bsbb ʔlʕnssr ‘ʔu: ʔllu:n ‘ʔu: ʔlʒns ‘ʔu: ʔllɣh ‘ʔu: ʔldii:n ‘ʔu: ʔlrʔi: si:a:si:aa: wɣi:r si:a:si: ‘ʔu: ʔlʔssl ʔlu:tsni: ʔu: ʔla:ʒtma:ʕi: ‘ʔu: ʔlθru:h ‘ʔu: ʔlmu:ld ‘ʔu: ʔi:i wdsʕ ʔxr. wfdsla:a ʕn ðlk la: jʒu:z ʔltmi:i:zu ʕli: ʔsa:s ʔlu:dsʕ ʔlsi:a:si: ʔu: ʔlqa:nu:ni: ʔu: ʔldu:li: llbld ʔu: ʔlʔqli:m ʔlði: jntmi: ʔli:h ʔlʃxss ‘su:a:ʔ ʔka:n mstqla:a ʔu: mu:dsu:ʕaa: tħt ʔlu:ssa:i:h ʔu: ɣi:r mtmtiʕ ba:lħkm ʔlða:ti: ʔm xa:dsʕaa: lʔi:i qi:d ʔxr ʕla: si:a:dth.

4. Conclusion

In summary, the website converts Arabic script into Buckwalter Arabic Transliteration, yielding a consistent representation in Latin letters. When users pick the “Buckwalter/IPA” tool, a table with three columns appears: “Arabic Script” for input, “Buckwalter Transliteration” for output text, and “IPA” for pronunciation. This tool is highly useful for linguistic study, natural language processing, and applications requiring a Latinized version of Arabic text. It ensures precision while preserving linguistic nuances. Overall, the site simplifies the process of dealing with Arabic language data while meeting a wide range of professional and academic requirements.

The Buckwalter Transcription System, developed by Tim Buckwalter’s brilliant intellect, is important in computational linguistics and Arabic language processing. It serve as an indispensable digital conduit, seamlessly translating Arabic text into ASCII characters. Over time, it has evolved into a cornerstone for linguistic exploration, providing scholars with a framework to navigate the intricacies of this and complex language. Our in-depth investigation explored the intricacies of the Buckwalter Transcription System. We delved into its comprehensive functionality, scrutinized its strengths and weaknesses, and analyzed the governing rules that underpin this transcription system. This journey illuminated the underlying principles driving its operation,

revealing the inner workings of this transformative linguistic tool.

At the culmination of our research, we have developed a user-friendly website, enabling users to convert Arabic script into Buckwalter Arabic Transliteration effortlessly. This endeavor exemplifies the system's versatility and adaptability to contemporary technological demands. The Buckwalter Transcription System transcends its status as a mere linguistic instrument; it testifies to the transformative potential of technology in surmounting linguistic barriers and unveiling the enigmatic facets of intricate languages like Arabic. Our anticipations for this work extend to advancing Arabic language processing, by making it more accessible, efficient, and user-friendly, thereby benefiting linguists and academics in their pursuit of linguistic understanding and exploration.

Author Contributions

Conceptualization, I.A.A. and T.A.; methodology, I.A.A. and T.A.; software, I.A.A. and T.A.; validation, I.A.A. and T.A.; formal analysis, I.A.A. and T.A.; investigation, I.A.A. and T.A.; resources, I.A.A. and T.A.; data curation, I.A.A. and T.A.; writing—original draft preparation, I.A.A. and T.A.; writing—review and editing, I.A.A. and T.A.; visualization, I.A.A. and T.A.; supervision, I.A.A. and T.A.; project administration, I.A.A. and T.A.; funding acquisition, I.A.A. and T.A. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2502.

Institutional Review Board Statement

Not applicable: this study did not involve human or animal subjects and therefore did not require ethical approval.

Informed Consent Statement

Not applicable: this study did not involve human participants and therefore did not require informed consent.

Data Availability Statement

No primary data were collected or analyzed in the present study; therefore, no datasets are available for sharing.

Acknowledgments

No additional acknowledgments are applicable for this study. All relevant contributions and support have been fully disclosed in the author contributions and funding sections.

Conflicts of Interest

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Habash, N., Rambow, O., 2005. Arabic, Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. In: Knight K, Ng HT, Oflazer K, (eds.). 43rd Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference. Association for Computational Linguistics: New Brunswick, NJ, USA. pp. 573–580.

[2] Diab M., Hacioglu K., Jurafsky D., Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, Boston, MA, USA, (2–7 May 2004); pp. 149–152.

[3] Buckwalter, T., 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania: Philadelphia, PA, USA. DOI: https://doi.org/10.35111/7vzm-mb15

[4] Buckwalter, T., 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium: Philadelphia, PA, USA. DOI: https://doi.org/10.35111/050q-5r95

[5] Al-Subaihin, A., Atwell, E., 2012. A Study of The Accuracy and Utility of Arabic Stemmers. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, (21–27 May 2012); pp. 619–625.

[6] Habash, N., Soudi, A., 2007. Buckwalter T. On Arabic Transliteration. In: Soudi, A., Van Den Bosch, A., Neumann, G., (eds.). Arabic Computational Morphology: Knowledge-Based and Empirical Methods. Springer: Dordrecht,NL, USA. pp. 15–22. DOI: https:

//doi.org/10.1007/978-1-4020-6046-5_2

[7] Hymes, D.H., 1982. Toward Linguistic Competence. University of Pennsylvania, Graduate School of Education: Philadelphia, PA, USA. pp. 9–23.

[8] Maamouri, M., GraffD, Jin, H., Cieri, C., et al., 2004. Dialectal Arabic Orthography-Based Transcription. Paper presented at: EARS RT-04 Workshop: Palisades, NY, USA.

[9] Alsadhan, N., 2025. A novel dialect-aware framework for the classification of arabic dialects and emotions. J Comput Sci. 21(1), 88–95.

[10] Abushaala, S., Elsheh, M., 2022. A comparative study on various deep learning techniques for arabic nlp syntactic tasks. Int J Comput Trends Technol. 70(1), 1–3.

[11] Fadel, A., Tuffaha, I., 2019. Al-Ayyoub M. Arabic Text Diacritization Using Deep Neural Networks. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia (1–3 May 2019); pp.1–7.

[12] ElSabagh, A.A., Azab, S.S., Hefny, H.A., 2025. A comprehensive survey on Arabic text augmentation: approaches, challenges, and applications. Neural Computing and Applications. 37, 7015–7048.

[13] Albahli, S., 2025. An advanced natural language processing framework for arabic named entity recognition: a novel approach to handling morphological richness and nested entities. Applied Sciences. 15(6), 3073.

[14] Wibawa, A.P., Kurniawan, F., et al., 2024. Advancements in natural language processing: Implications, challenges, and future directions. Telematics and Informatics Reports. 16, 100173. DOI: https://doi.org/10.1016/j.teler.2024.100173

[15] Gorgis, D.T., 2010. Translaiterating Arabic: The Nuisances of Conversion between Romanization and Transcripting Schemes. In: Izwaini S. (ed.). Romanization of Arabic Names: Proceedings of the International Symposium on Arabic Transliteration Standard: Challenges and Solutions, Abu Dhabi, UAE; (15-16 December 2009); pp. 20–21.

[16] Abdul-Mageed, M., Diab, M., 2014. Kübler S. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language. 28(1), 20–37.

[17] Nawar, M.N., 2014. Improving Arabic Tokenization and Pos Tagging Using Morphological Analyzer. In Advanced Machine Learning Technologies and Applications: Second International Conference, AMLTA 2014, (28–30 November 2014); Cham: Springer International Publishing: Cairo, Egypt. pp. 46–53.

[18] Habash, N., Rambow, O., Roth, R., 2010. MADA+TOKAN Manual. A toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. Columbia University: New York, NY, USA. DOI: https://doi.org/10.7916/d86d60bs

[19] Farber, B., Freitag, D., Habash, N., et al., 2008. Improving NER in Arabic Using a Morphological Tagger. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, (26–30 May 2008).

[20] Habash, N., Roth, R., 2008.Identification of Naturally Occurring Numerical Expressions in Arabic. Lang Resour Eval. 42(3), 333–336.

[21] Hajic, J., 2000. Morphological Tagging: Data Vs. Dictionaries. In 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, USA, (29 April–4 May 2000); Association for Computational Linguistics: Stroudsburg, PA, USA.

[22] Broselow, E., McCarthy, J., Eid, M., (eds.), 1992. Perspectives on Arabic linguistics IV: Papers from the Fourth Annual Symposium on Arabic Linguistics. John Benjamins : Amsterdam, Netherlands.

[23] Al-Sughaiyer, I.A., Al-Kharashi, I.A., 2004. Arabic morphological analysis techniques: A comprehensive survey. Journal of the Association for Information Science and Technology. 55(3), 189–213. DOI: https://doi.org/10.1002/asi.10368

[24] Al-Shenaifi, N., Azmi, A.M., Hosny, M., 2024. Advancing AI-Driven Linguistic Analysis: Developing and Annotating Comprehensive Arabic Dialect Corpora for Gulf Countries and Saudi Arabia. Mathematics. 12(19), 3120.

[25] Faheem, M,A,, Wassif, K.T., Bayomi, H., et al., 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of Egyptian dialect to modern standard Arabic translation. Scientific Reports. 14(1), 2265.

[26] Omar, L.I., Salih, A.A., 2024. Systematic review of english/arabic machine translation postediting: Implications for AI application in translation research and pedagogy. Informatics. 11(2), 23.

[27] Beesley, K.R., 1998. Arabic morphology using only finite-state operations. In: Rosner, M.(ed.). Computational Approaches to Semitic Languages: Proceedings of the Workshop. Association for Computational Linguistics: Montreal, QC, Canada. pp. 50–57. DOI: https://doi.org/10.3115/1621753.1621763

[28] Ahmed, N., Saha, A.K., Al Noman, M.A., et L., 2024. Deep learning-based natural language processing in human-agent interaction: Applications, advancements and challenges. Natural Language Processing Journal. 28: 100112.

[29] Habash, N., 2007. Arabic Morphological Representations for Machine Translation. In: Habash, N., (ed.). Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer: Dordrecht, NL, SUA. pp. 263–85. DOI: https://doi.org/10.1007/978-1-4020-6046-5_14

[30] Alluhaybi, I., Witzel, J., 2020. Letter connectedness and Arabic visual word recognition. Quarterly journal

of experimental psychology. 73(10), 1660–1674.

[31] Witzel, J., Cornelius, S., Witzel, N., et al., 2015. Testing the Viability of WebDMDX for Masked Priming Experiments. In: Jarema, G., Libben, G., (eds.). Phonological and Phonetic Considerations of Lexical Processing. John Benjamins:Amsterdam, Netherlands. pp. 169–98.

[32] Saleh, H., AlMohimeed, A., Hassan, R., et al., 2025. Advancing arabic dialect detection with hybrid stacked transformer models. Frontiers in Human Neuroscience. 19, 1498297. DOI: https://doi.org/10.3389/fnhum.2025.1498297

[33] Darwish, K., 2002. Building a Shallow Arabic Morphological Analyzer in One Day. Association for Computational Linguistics: Stroudsburg, PA, USA. pp. 1–8. DOI: https://doi.org/10.3115/1118637.1118643

[34] Khoja. S., 2001. APT: Arabic Part-of-Speech Tagger. Association for Computational Linguistics: Stroudsburg, PA, USA. pp. 20–25.

[35] Wikimedia., 2025. Buckwalter Transliteration. Available from: https://en.wikipedia.org/wiki/Buckwalter_transliteration (cited 5 June 2025).