

# English-Arabic Phonetic Dataset construction

Zaid Rajih Mohammed<sup>1\*</sup>, and Ahmed H. Aliwy<sup>2</sup>

<sup>1</sup>Jaber bin Hayyan University of Medicaland Pharmaceutical Sciences

<sup>2</sup>Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq

**Abstract.** In the field of natural language processing, the effectiveness of a semantic similarity task is significantly influenced by the presence of an extensive corpus. While numerous monolingual corpora exist, predominantly in English, the availability of multilingual resources remains quite restricted. In this study, we present a semi-automated framework designed for generating a multilingual phonetic English-Arabic corpus, specifically tailored for application in multilingual phonetically and semantic similarity tasks. The proposed model consists of four phases: data gathering, preprocessing and translation, extraction IPA representation, and manual correction. Four datasets were used one of them was constructed from many sources. A manual correction was used at all the levels of the system to produce a golden standard dataset. The final dataset was in the form (English Word, English Phonetic, equivalent Arabic Word, and Arabic Phonetic). Also, a deep learning approach was used for extracting International Phonetic Alphabet (IPA) phonetic representation where the results for 13400 samples show that the Phonetic Error Rate (PER) and accuracy were 11.96% and 88.04 % respectively which are good results for producing IPA representation for unknown English and Arabic names.

## 1 Introduction

Phonetic transcription is a form generated by the conversion of grapheme sequence (spelling of the word) into phoneme sequence (phonetic form), for example, the word “Computer” is phonetically transcribed as /kəmˈpjʊtər/. Phonetic transcription is an important element in speech recognition, text-to-speech applications, end-to-end speech translation systems, and the development of speech databases [1]. Extracting the phonetic representation of words is a difficult task for many complex languages. These difficulties are increased when we try to map a phonetic presentation of a word in one language into a phonetic representation in another language such as proper-name transliteration that can be used in many applications and tasks.

Many class types in natural language such as names, verbs, adjectives, etc. are increasing daily result of the rapid growth of technologies and inventions. These new words should pronounced accurately in other languages and hence they should be transcribed phonetically in these languages [2].

Most of the methods that map from phonetic representation in one language to phonetic representation in another language use rules-based, machine learning, or deep learning approaches, and these techniques require datasets for learning, testing, or evaluation. This dataset should have a specific structure as a source phonetic representation and target phonetic representation. Many languages suffer from missing such a dataset or existing a very small dataset that is not sufficient for machine learning algorithms, for example, the English-Arabic phonetic dataset does not exist according to our knowledge [11].

Therefore, we are trying to construct a bilingual phonetic dataset for English and Arabic languages to be used as a standard dataset for many applications and tasks.

---

\* Corresponding author.: [zaid.rajah@jmu.edu.iq](mailto:zaid.rajah@jmu.edu.iq)

The remainder of the paper is structured as follows. Section 2 reviews similar kinds of work for different world languages. The proposed model is presented in Section 3. The experiment of our model and the discussion of the result of the dataset bilingual are shown in Section 4, while the conclusion is shown in Section 5.

## 2 Related works

Most of the available datasets are represented as monolingual-phonetic works that used private and public phonetics, in this section a survey for the related works will be presented

Alshuwaier and Areshey [3] used a private 100 samples dataset based on Google and Bing translation, for testing a translation of English names to Arabic based on manual rules. It has very little and hence cannot be used for learning any ML model. They get 18.5% and 82.1% for Google and Bing respectively. Toma et. al [4] introduces a phonetic dictionary for Romanian, It contains over 70,000 words, and their manually performed phonetic transcription. And building a grapheme-to-phoneme converter based on decision trees. Beáta [5] evaluated four different sequence-to-sequence deep neural network architectures aimed to jointly solve the tasks of phonetic transcription, lexical stress assignment, and syllabification, and he used two datasets MaRePhor and CMUdict.

Lee et.al [6] introduced WikiPron, a tool for extracting pronunciation data from Wiktionary, a collaborative multilingual online dictionary. This tool did not provide multilingual for the same word. Yolchuyeva et. al [9] introduced convolutional neural networks (CNNs) for the conversion of graphemes to phonemes in the context of grapheme-to-phoneme (G2P) conversion. The paper outlines five distinct models designed for the G2P task and compares their outcomes with previously documented state-of-the-art research. Their model is constructed based on the seq2seq architecture, and in the fourth and fifth models, they incorporate CNNs with residual connections. However, this research did not address the handling of bilinguals.

Loots and Niesler [8] used data-driven to conversion grapheme to phoneme and analysis pronunciations. This research delved into the characteristics and relationships among three distinct English accents (American, British, and South African) as represented by many pronunciation dictionaries such as (CMU-DICT). The internal consistency of these dictionaries was assessed through G2P algorithms. Additionally, the G2P algorithm underwent modification to facilitate the conversion of pronunciations from one accent to another. Engelhart et. al in [13] They presents a transfer learning method for cross-dialect learning in English, utilizing a neural G2P model with a Transformer architecture. The approach achieves impressive accuracy, particularly for dialects with limited available data. However, the study does not explore the application of addressing multilingual challenges.

Our work emphasizes on construction of a dataset for proper names and loanwords between English and Arabic languages with dual-phonetic transcription for the same word.

## 3 Proposed model

According to our goal for the construction of an English-Arabic phonetic transcribed dataset, an international phonetic representation should be selected. For this task International Phonetic Alphabet (IPA) can be used which is the most famous standard phonetic alphabet. The proposed model of dataset construction consists of four phases: data gathering, preprocessing and translation, extraction IPA representation, and manual correction.

The output will be a bilingual English–Arabic phonetic representation in the form (English Word, English Phonetic, equivalent Arabic Word, and Arabic Phonetic). Figure (1) the main diagram of the proposed model.

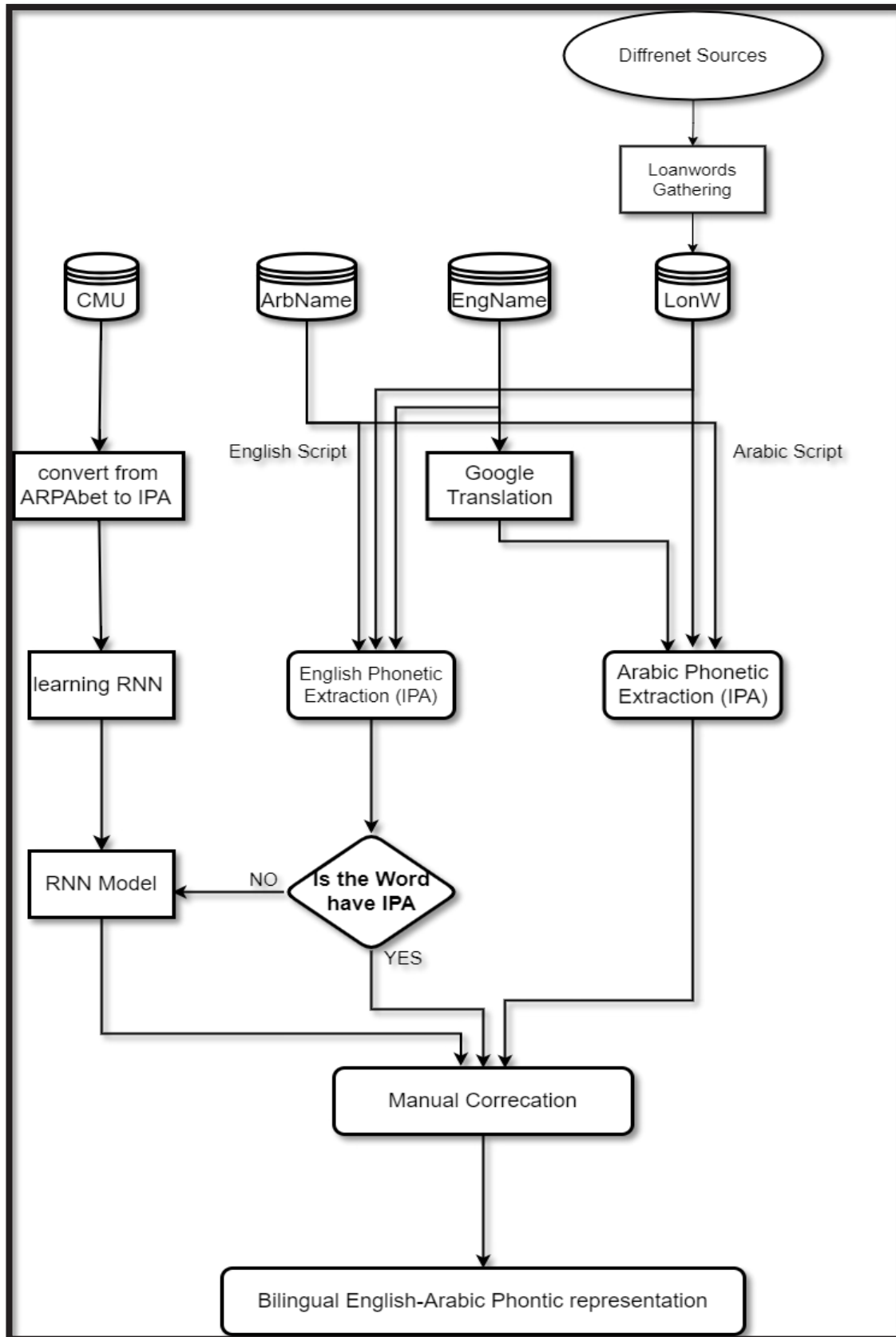


Fig.1. shows the main diagram of the proposed model

## 4 The used Datasets

Four types of datasets were used which are the CMU Pronouncing Dictionary, Database Hub, Hamari Web, and loanword. These datasets are different in the used languages, phonetic representation, number of words, and scripts. The first three datasets were publically available for the researchers in different sites while the Loanword dataset was constructed from different sources by the authors. Table (1) show the statistics of different resources.

**Table (1)** the statistics of different resources.

No	Data	Name	Script Arabic	Script English	Number of words	English Phonetic representation	Arabic Phonetic representation
1	CMU dictionary	CMU	No	Yes	132k	Yes	No
2	Database Hub	EngName	No	Yes	250k	No	No
3	Hamari Web	ArbName	Yes	Yes	23k	No	No
4	Loanwords	LonW	Yes	Yes	0.5k	No	No

CMU (The pronouncing dictionary) is an open source, from Carnegie Melon University, that contains over 134000 English words and their pronunciations. The used phonetic representation is ARAbet representation and hence all the phonetic representation should be transformed into IPA. EngName is a list of English names only without any translation or phonetic representation that available on the database hub. ArbName is list of Arabic names without phonetic representation that are available on Hamari web. LonW, as was mentioned previously, was constructed from different sources in the format Arabic-English or English-Arabic without phonetic representation. Table () shows information about the collected data.

## 5 Preprocessing and translation

At this stage, a set of steps were taken, such as deleting of the duplicated names and words, removing special letters, removing diacritics, Normalize Arabic characters to a standardized form and letters unification. EngName has monolingual representation as English scripts only therefore, it should be translated into Arabic language scripts. Google translation is used for this task because it is freely available and gives an acceptable result. All the translated names are corrected manually, for example the English word “Ray” is translated wrongly into “ray”-“الاشعة”.

## 6 Extraction IPA representation

The used datasets contains English and/or Arabic words with/without phonetic representation. Therefore three separate steps can be applied

- Extracting phonetic representation for Arabic scripts that has no phonetic representation.
- Extracting phonetic representation for English scripts that has no phonetic representation.
- Unification of phonetic representation for English and Arabic scripts.

For this task International Phonetic Alphabet (IPA) can be used which is the most famous standard phonetic alphabets.

Extracting phonetic representation, for Arabic scripts that has no phonetic representation, can be done using Epitran library which is used to find the phonetic representation of many languages, including Arabic. For all the Arabic scripts, a phonetic representation will be extracted therefor there is not any out of vocabulary (OOV) case. The output will be in IPA representation. For English

Extracting phonetic representation, for English scripts that has no phonetic representation, can be done using eng\_to\_ipa library which is used to find the phonetic representation of English only. It is used, in this research, because it is more focused choice and more straightforward and specific for English. Some of English scripts may have no phonetic representation because eng\_to\_ipa library depend on CMU dictionary and any script that out of this vocabulary will not has IPA representation. In this case, a learned model should be used for this task as will explained in next sections. For this task, a deep learning model RNN based on LSTM that constructed by [6], is used to find the phonetic representation (IPA) of a script. Our model was learned from CMU dataset. Figure (2) shows LSTM phonetic representation extraction model.

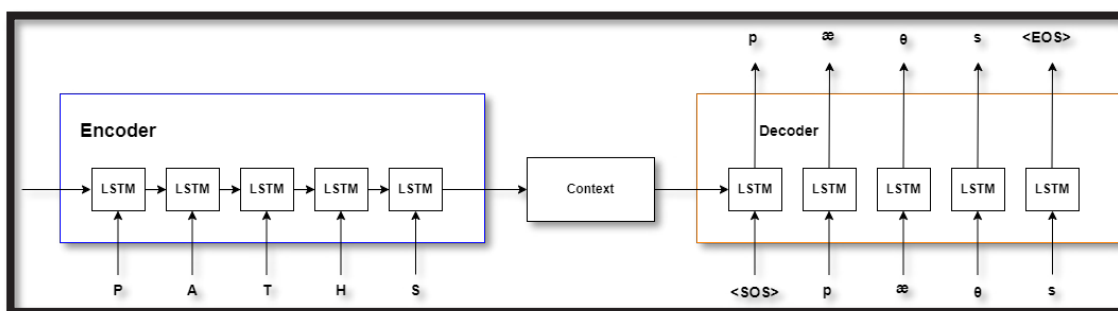


Fig.2. LSTM model for phonetic representation extraction.

Unification of phonetic representation, for English and Arabic scripts, is done by selecting IPA representation and transforming the none IPA into IPA. CMU dataset in the form of ARBAbet representation therefore it should be transformed into IPA. A direct mapping between ARBAbet representation symbols and IPA representation symbols was used.

## 7 Manual correction

Manual correction was adopted in more than one stage of creating the dataset, from the data gathering stage, which required manual intervention in collecting loanwords, as well as in the translation and pre-processing stage, where knowledge of the true translation of the name was required. At this stage, manual modification was relied upon because the outcomes of deep learning, depend on finding the correct pronunciation of words written in English, but most Arabic nouns are written in English differently from one person to another, so it requires manual intervention to choose the pronunciation closest to the Arabic pronunciation, and this was done based on the opinion of experts.

## 8 Experiment results and Evaluation

All tests were conducted using the Python language on a computer with 8 GB RAM Intel Core i7 processor using many libraries. Four types of datasets were used one of them (LonW) was constructed by us from many sources. Figures (3), (4), (5) and (6) show samples of CMU, Database Hub, Hamari Web, and lonW datasets.

**Table (2)** samples of CMU.

English Word	ARAPbet
dena	D IY N AH
denapoli	D IH N AA P AH L IY
denard	D IH N AA R D
denardo	D IH N AA R D OW
denarii	D IH N AE R IY
denarius	D IH N AE R IY AH S
denaro	D IH N AA R OW
denatale	D IH N AA T AA L IY
denationalization	D IY N AE SH AH N AH L IH Z EY SH AH N
denationalizations	D IY N AE SH AH N AH L IH Z EY SH AH N Z
denationalize	D IH N AE SH AH N AH L AY Z
denationalized	D IH N AE SH AH N AH L AY Z D
denationalizing	D IH N AE SH AH N AH L AY Z IH NG
denature	D IH N EY CH ER

List of Names Dataset / Complete\_List\_Names

Complete dataset with more than 250k male and female names.

Data
Schema

Name	Gender
Will	male
James	male
Samuel	male
John	male

**Fig. 3** samples of Database Hub.

Abrar	Virtuous, Pious, Great Man	ابرار
Muslim	Submitting Oneself To Allah	مسلم
Hasan	Beautiful, Gentle, Handsome Man, Grandson Of Prophet	حسن
Sameer	Jovial, Beneficial, Entertaining Companion, Good Friend	ثمير
Aman	Safety, Protection, Peace	امان
Umair	Life, Long-Lived, Intelligent Man	عمير
Ahmad	Praiseworthy, Commendable, Noble, Admirable	احمد
Shoaib	Who Shows The Right Path, A Guide, Name Of Prophet	شعيب

Fig.4 a samples of Hamari Web.

Table 3 a samples of lonW datasets.

English word	Original
apricot	البيوق
arsenal	دار الصناعة
artichoke	ارض شوك
assassin	الحشاشي
jar	جرة

figure (7) shows the LonW dataset after the manual corrections. Figure (8) show the Google translation of English names for Database Hub with/without manual correction.

Table 4. the LonW dataset after the manual corrections.

English word	Original	Manual Corrections
apricot	البيوق	برقوق
arsenal	دار الصناعة	دار صناعة
artichoke	ارض شوك	ارضيشوك
assassin	الحشاشي	حشاشي
jar	جرة	جرة

**Table 5** the Google translation of English names for Database Hub with/without manual correction.

English Name	Google Translation	Manual Corrections
John	جون	جون
William	وليام	وليام
James	جوامع	جمس
Charles	تشارلز	تشارلز
George	جورج	جورج
Frank	رصیح	نراذک
Joseph	جوزيف	جوزيف
Thomas	توماس	توماس
Henry	بيہ	بيہ
Robert	روبرت	روبرت
Edward	إدوارد	إدوارد
Harry	ھاري	ھاري
Walter	والب	والب
Arthur	آرثر	آرثر

For Extraction of IPA phonetic representation step, figure (9) show Arabic IPA phonetic representation, figure (10) show English IPA phonetic representation, and figure (11) show mapping between ARBAbet and English IPA phonetic representation.

**Table 6.** Arabic IPA phonetic representation

Arabic Word	Arabic Phonetic
عبد العزيز	/ʔ b d a: l ʔ z i: z/
محمد	/ m f i m d /
حرب	/ f i r b /
حسن	/ f i s n /
هاشم	/ h a: ʃ m /
قرآن	/ q r b a: n /
راغب	/ r a: ɣ b /
نجار	/ n dʒ a: r /

**Table (7) English IPA phonetic representation**

English Word	English Phonetic
Gust	/gəst/
Vainest	/vaɪnseɪnt/
Silvestre	/sɪlvɛstər/
Fredrik	/frɛdrɪk/
Francisco	/frænsɪskoʊ/
Erick	/ɛrɪk/
Goldin	/goʊldən/

**Table (8) mapping between ARBAbet and IPA.**

English Word	ARAPbet	IPA
dena	D IY N AH	dɪnə
denapoli	D IH N AA P AH L IY	dɪnəpəli
denard	D IH N AA R D	dɪnɑrd
denardo	D IH N AA R D OW	dɪnɑrdəʊ
denarii	D IH N AE R IY	dɪnəri
denarius	D IH N AE R IY AH S	dɪnəriəs
denaro	D IH N AA R OW	dɪnɑrəʊ
denatale	D IH N AA T AA L IY	dɪnatəli
denationalization	D IY N AE SH AH N AH L IH Z EY SH AH N	dɪnæʃənəlaɪzɪʃən
denationalizations	D IY N AE SH AH N AH L IH Z EY SH AH N Z	dɪnæʃənəlaɪzɪʃənz
denationalize	D IH N AE SH AH N AH L AY Z	dɪnæʃənəlaɪz
denationalized	D IH N AE SH AH N AH L AY Z D	dɪnæʃənəlaɪzɪd
denationalizing	D IH N AE SH AH N AH L AY Z IH NG	dɪnæʃənəlaɪzɪŋ
denature	D IH N EY CH ER	dɪneɪtʃər

For evaluation of deep learning model for extracting the IPA to OOV names, the CMU dictionary was used. It was split into 85% for the training set, 5% for validation, and 10% test set, and our model was evaluated by computing PER and accuracy. PER is defined as the number of insertions, deletions, and substitutions divided by the number of true phonemes. For 13400 samples, the PER and accuracy were 11.96% and 88.04 % respectively which is good results for producing IPA representation for unknown English Names.

The final step, the core of our research, is constructing the final Bilingual English- Arabic dataset with dual IPA phonetic representation. Figure (12) shows a sample of the final dataset. Table (2) shows the statistics of the final dataset.

**Table 9.** the statistics of the final dataset.

No	Data	Script Arabic	Script English	Size	English Phonetic representation	Arabic Phonetic representation
1	English names	No	Yes	3857	Yes	Yes
2	Arabic names	Yes	Yes	2408	Yes	Yes
3	Loanwords	Yes	Yes	235	Yes	Yes
SUM	-	-	-	6500	Yes	Yes

**Table 10.** sample of the final dataset.

English Word	English Phonetic	Arabic Phonetic	Arabic Word
Gust	/gəst/	/yust/	غوست
Vainest	/vɪnsɪnt/	/fnst/	فنتست
Silvestre	/sɪlvɛstər/	/sɪlfsɪr/	ب:سلفس
Kohl	/koʊl/	/kfil/	كحل
Lemon	/lɛmən/	/li:mun/	ليمون
magazine	/mægəzɪn/	/mxazn/	مخازن
Makkah	/mekə/	/mkh/	مكة
Safari	/səfɑrɪ/	/sfr/	سفر
Saffron	/sæfrən/	/zʃfrɑ:n/	زعفران
Spinach	/spɪnəʃ/	/sbɑ:nx/	سبانخ
Saltant	/sɔltən/	/slʔɑ:n/	سلطان
Syrup	/sɪrəp/	/frɑ:b/	رشاب
Talisman	/tælɪsmən/	/ʔlsm/	طلسم
Abdulaziz	/æbduləzɪz/	/ʃbdɑ:lʃzɪ:z/	عبد العزيز
Mohammad	/məhɑməd/	/mfɪmd/	محمد
Harb	/hɑrb/	/firb/	حرب
Hasan	/hɑsɑn/	/fisn/	حسن
Hashim	/hæʃɪm/	/hɑ:ʃm/	هاشم

When using the proposed model to find the phonetic representation of Arabic words that translated to English and it is OOV, we observed that the results were limited or unsatisfactory compared to the original English that OOV.

Also, one of the challenges facing researchers and individuals interested in the Arabic language is that Arabic names, when transliterated into English, are written in various forms depending on the individual's pronunciation. This poses a problem, as highlighted in some studies such as [7, 12]. For example, the Arabic name "تحسين" (Tahseen) is noted to have multiple transliterations such as "Tahseen" and "Tahsin".

## 9 Conclusion

In this paper, we presented a bilingual dictionary that can be used in phonetic mapping between the two languages (English and Arabic). As mentioned previously, the proposed model consists of four phases: data gathering, preprocessing and translation, extraction IPA representation, and manual correction. Four datasets were used one of them was constructed from many sources. Manual correction was used at all levels of the system to produce a golden standard dataset. The final dataset was in the form (English Word, English Phonetic, equivalent Arabic Word, and Arabic Phonetic).

According to the results, producing IPA for unknown words is a difficult task for English and Arabic words or names but it can be achieved using a good learned model. The absence of standard translation for the names causes problems in learning and extracting IPA phonetic representation. Also using `epitran` and `eng_to_ipa` libraries for producing IPA phonetic representation has many errors for unknown words (OOV). The deep learning model was a good choice for extracting IPA phonetic representation for both the Arabic and English languages.

In future works, we suggest extending the constructed dataset with another language such as Persian or any other language. Also, other deep learning techniques can be used to compare the results of producing IPA phonetic representation.

## References

1. T. Yuanhe, L. Renze, P. Xiangyu W. Lianxi J. Shengyi and S. Yan ,” Improving English-Arabic Transliteration with Phonemic Memories”, Findings of the Association for Computational Linguistics, pages 3262--3272, 2022.
2. . K. Nahar, . H. Al-Muhtaseb, . W. Al-Khatib, . M. Elshafei and . M. Alghamdi, "Arabic Phonemes Transcription using Data Driven Approach.," International Arab Journal of Information Technology (IAJIT), vol. 12, 2015.
3. F. Alshuwaier and A. Areshey, “Translating English Names to Arabic Using Phonotactic Rules”, 25<sup>th</sup> Pacific Asia Conference on Language, Information and Computation, pages 485–492, 2011.
4. Ş.-A. Toma, A. Stan, M.-L. Pura and T. Bârsan, "MaRePhoR—An open access machine-readable phonetic dictionary for Romanian," in 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2017.
5. B. Lőrincz, "Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks," *Procedia Computer Science*, vol. 176, pp. 108–117, 2020.
6. J. L. Lee, L. F. Ashby, M. E. Garza, . Y. Lee-Sikka, . S. Miller, . A. Wong, A. D. McCarthy and . K. Gorman, "Massively multilingual pronunciation modeling with WikiPron," in Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020.
7. . K. Rao, . F. Peng, . H. Sak and . F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.
8. . L. Loots and . T. Niesler, "Data-driven phonetic comparison and conversion between South African, British and American English pronunciations," in Tenth Annual Conference of the International Speech Communication Association, 2009.
9. S. Yolchuyeva, G. N'emeth and B. Gyires-T'oth, "Grapheme-to-phoneme conversion with convolutional neural networks," *Applied Sciences*, vol. 9, p. 1143, 2019.
10. A. P. Saucedo, A. S. Sepúlveda and D. F. Gómez Cajas, "Phoneme Recognition System Using Articulatory-Type Information," *Tecciencia*, vol. 10, pp. 11--14, 2015.
11. . A. Freeman, . S. Condon and . C. Ackerman, "Cross linguistic name matching in English and Arabic," in Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006.

12. . K. Yao and . G. Zweig, "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion," arXiv preprint arXiv:1506.00196, 2015.
13. . E. Engelhart, . M. Elyasi and . G. Bharaj, "Grapheme-to-phoneme transformer model for transfer learning dialects," arXiv preprint arXiv:2104.04091, 2021