

## Evaluation of domain sulfur industry for DIA translator using bilingual evaluation understudy method

Huda Mohammed Lateef<sup>1</sup>, Ahmad Muter Awaad<sup>2</sup>, Diadeen Ali Hameed<sup>3</sup>, Ghanim Thiab Hasa<sup>3</sup>, Tahseen Ameen Faisal<sup>4</sup>

<sup>1</sup>Electronic Computer Center, University of Fallujah, Anbar, Iraq

<sup>2</sup>Ministry of Education, Anbar Directorate of Education, Anbar, Iraq

<sup>3</sup>Department of Electrical Engineering, College of Engineering Alshirqat, University of Tikrit, Tikrit, Iraq

<sup>4</sup>Department of English Language, College of Basic Education- Shirqat, Tikrit University, Tikrit, Iraq

### Article Info

#### Article history:

Received Jul 30, 2022

Revised Dec 21, 2022

Accepted Apr 12, 2023

#### Keywords:

Bilingual evaluation understudy method

Evaluation English-Arabic machine translation

Evaluation machine translation

Machine translation

Neural approach machine translation

### ABSTRACT

Evaluation is important part of our system development cycle; it also contributes to improving new machine translation (MT) technology optimum via comparing them with the traditional systems available to determine the weaknesses and the effectiveness to be improved in the proposed MT system. This work aiming to make a study that evaluate the performance and effectiveness of the domain sulfur industry (DSI) for English-Arabic DIA translator quality. The recent study has conducted evaluating by making a comparison between this programme with the prominent Google translator through applying a rendering of 1,200 English sentences in bilingual evaluation understudy (BLUE) method. The obtain results show that the efficiency of Google translator is about 30.325%, while DIA translator efficiency in domain sulfur industry is about 73.325% and it's more effective and give a better translation accuracy. The BLUE method efficiency is about (90.478%) compared with the human expert evaluator.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



### Corresponding Author:

Diadeen Ali Hameed

Department of Electrical Engineering, College of Engineering Alshirqat, Tikrit University  
Tikrit, Iraq

Email: diaa@tu.edu.iq

## 1. INTRODUCTION

Evaluating the systems of machine translation (MT) is a considerable field of researches to optimize the effectiveness of technologies of MT improvement cycle [1]. Evaluation refers to estimate or examine the validity of a particular thing. Anytime a particular novel technology is to be under development. it again requires updated testing or assessment on particular bases. Similarly, the requirement forevaluating the MT arise [2]. Given the great development in the system of MT field, as well as the prominence of the requirement for great speed and extremely high degree of accuracy in the information convenience interchangeable between two or more languages [3], human evaluation consumes more time in addition of being expensive and thus inappropriate to be used repeatedly during research or develop MT system engines [4].

Evaluating of the system of MT and the system of MT itself are of the same importance, tackling issues concerning the interpretation of linguistic item precisely, fluently, and in an acceptable acceptable say [5], and then attestattesting an MT algorithm. During the last dozens of years, it has been used a huge number of metrics for evaluating the quality of MT, on the ground of a variety of similar standards that are proposed to be an independent tongue and not targeted a certain natural language. The majority of them are relied on comparing between the automatic translation and that of direct reference [6].

Any machine translation accuracy is normally estimated through making a comparison between the outcomes of it with those of expertise human judgements [7]. The recent study has been conducted on the performance-based method. BiLingual evaluation understudy (BLEU) which is introduced by Papineni *et al.* [2] is a method used to evaluate MT systems, which is supposed to be autonomous language independent and greatly based upon the human assessment. BLEU is highly constructed on an essential notion for determining the goodness of a particular MT programme. It could be made briefly by the proximity of the proposed outcome of the MT scheme with indication to a translated text done by an (experienced human) translation of the text itself [8].

The proximity of the selected translation to the referred one is decided by a mutated n-gram accuracy when  $n=\{1, 2, 3, 4\}$  [9]. The mutated n-gram accuracy is the essential standard that BLEU apply to differentiate among well done and weak selected translations [10], as this standard is centred on calculating the amount of highly occurred words in the selected translation as well as the referred rendering, followed by dividing the amount of the highly occurred words by the gross amount of words in the selected rendering [11]. The mutated n-gram accuracy determines selected linguistic structures as being shorter than those of referred opposite parts [12] in addition, this n-gram determines selected linguistic structures which have over generated correct word forms.

English-to-Arabic MT has been an annoying and exciting research subject for a high number of researchers in the domain of processing standard Arabic language. A significant amount of attempts had been conducted for performing or improve MT from Arab language into many different ones [13]. This research concentrates on the assessment of the performance of the English-Arabic DIA MT software and the production of Google Translate. The purpose behind the recent research is to get an estimation for the conduction of DIA programme in comparison with that of Google Translate by dealing with a variety of text types directed from English into Arab language, as well as the quality of being acceptable acceptable and usable for the end-users. The adequacy scores [7], [14] and fluency scores are the main tests used to assess the quality of the translation [15].

## 2. METHOD

The recent research adopts the BLEU method [16], [17] for evaluating DSI for English-Arabic DIA translator and the Google translator. The evaluation conducted automatically just supple a way that compare the output texts with that of human references without absolutely measuring the goodness of the translation. Arab language uses variety of forms and arrangements for words, so as it could communicate any idea in various forms. Moreover, the so many dialects existed and the merit of being expressed in various forms is not necessarily similar concerning the two involved languages expectedly results in the probability of indicating so many meanings for only one sentence as it is Alqudsi *et al.* [18].

In these studies, the measurement of being intelligible is centred on a pair of characteristics s, i.e. being fluent as well as being adequate by using BLEU-score formula. It is resulted from the division of the brevity penalty (BP) by the geometric mean of altered n-gram accuracies. Therefore, we must begin by calculating the geometric mean of n-gram's altered accuracy. After that, the size of the candidate's text (c) and the duration of the effective reference corpus (r) must be calculated so as to be ready for calculating the BP. Then the closest human judgment score is determined. In (1) [2] demonstrates the way to generate a BP exponentially reduced (r/c):

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases} \quad (1)$$

In (2) shows the way of computing the final BLEU score:

$$BLUE = BP * \exp \sum_{n=1}^N w_n \log p_n \quad (2)$$

Whereas N equals 4, while regular weights  $w_n$  equals (1/N) [7].

The BLEU metric scores are ranging from 0 to 1 [2]; where the value (1) implies that the applicant text has fully matched the reference form, and the value (0) implies that the applicant text and the respective reference text are totally distinct. In view of the fact that the phrase is the fundamental unit in the translation process of the two programs assessed, it was selected as the fundamental test element. As a result of this research, only the output quality of sentences was assessed, the focus was on the preservation of meaning, which involves a comparison of meaning in output with that in the original [19].

Pre-processing data by separating any version into distinct n-gram dimensions, like the following: (uni)grams, (bi)grams, (tri)grams, and (tetra)grams. The accuracy of the DIA translator system and the Google translator was calculated for each of the four gram dimensions. Calculate a unified accuracy rating

for each of the four n-gram dimensions. These scores are then contrasted to decide which of them will get the highest version [20] (compare MT schemes: individual devices and system components are rated on the basis of how often they are considered to be superior than or equivalent to any other scheme). Algorithm that follows is applied to evaluate the translation as in the main steps: i) start; ii) input (source text); iii) input (two reference of target text); iv) translation source text by DIA translator; v) translation source text by Google translator; vi) automatic evaluating of DIA translator quality; vii) automatic evaluating of Google translator quality; viii) compare between DIA and Google output quality; ix) compare result quality by human expert evaluator; x) print (rank MT systems from best to worst); and xi) End.

The BLEU is quite a rough measure of translation performance [21], Figure 1 illustrates the main steps of the method and the way of extracting n-grams from English, Arabic, Arab language references of linguistic structures for calculating BLEU scores concerning the systems of MT of DIA translate plus Google translator. After that, the nearest human assessment could be judged. A variety significant factors can provide a contribution to the bilingual evaluation understudy (BLUE) grossness [22]: i) synonyms and paraphrases will only be used if they are in a collection of various reference types [23]; ii) word results are similarly weighted so that there is no extra punishment for missing content-bearing content [24]; and iii) the punishment for brevity is a stop-gap measure to compensate for the relatively severe issue of not being prepared to calculate recall [25]. Each of these mistakes leads to an enhanced number of inappropriately indistinguishable transmissions in the assessment. Since BLEU can theoretically assign equivalent scores to translations of manifestly distinct performance [26], it is logical that a greater BLEU rating is not possible.

Input	Source text				
	Reference 1,2,...,n				
Process	Google translate				
	DIA translate				
	BLUE method	1-gr.	2-gr.	3-gr.	4-gr.
	Google translate				
	DIA translate				
Output	Rank system from Best to Worst	1) - 2) -			

Figure 1. The structure of the main method steps

### 3. RESULTS AND DISCUSSION

We created software of automatic evaluation on Arabic MT quality (BLUE method) by applying Asp.net 2017 to execute this task. The quality evaluation of MT is illustrated by the quality evaluation of MT is illustrated by Table 1 shows, Adequacy: Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted? while Table 2 shows, Fluency: Is the output good fluent English? This involves both grammatical correctness and idiomatic word choices. Figure 2, which describes the main screen of the system of evaluating MT. Most proposed approaches for English-Arabic DIA translator have been tested on limited domain; sulfur industry. So, for evaluating the obtained outcomes of this scheme of evaluation, we selected a corpus of 1,200 phrases which are categorised under 4 criteria; terms, phrase, text with limited domain, and general text and then they were rendered into their counterparts in Arabic language by making use of each of the Babylon and Google translators.

The results obtained through using the BLUE method by the system of DIA programme as well as the application of Google translation, we have reached a conclusion of the the following:

- The analysis of using chemical symbols shows that Google data base doesn't include those symbols concerning the field of translating sulfur industry, on the contrary to DIA system which shows an integrated information of the input symbols because of being specialized informative system for translation in this field.
- The analysis of terminology test which is often not more than three words, also shows that Google data base hardly includes few terms compared with DIA system which was able to translate them.

- Testing specialized sulfur industry expressions shows that DIA system is highly better than Google.
  - Testing texts of no more than 50 words shows that DIA has the priority in showing the translated synonyms. While, both (DIA and Google) systems were equal concerning grammatical order of sentence constituents.
  - Testing common texts of no more than 50 words shows that Google has the priority in the translation because its data base is richer than that of DIA. Concerning texts other than the field of sulfur industry.
  - It has been generally observed that the Google Translate scheme has been normally noticed to be inferior in most applications as it compared with the system of DIA as indicated in Table 1.
  - Finally, by analysing human evaluation, and comparing the results by using BLEU method in the translations of both (DIA and Google) shows that BLEU method was of (89.875%) adequacy.
- Concerning the rate of the degree of accuracy of results for each phrase of the Google and BLEU methods corpus, DIA programme confirmed greater translation accuracy than that of Google Translate (73.325%) concerning DIA, while 30.325% concerning Google for the performed tests. The systems of MT are illustrated in both of Figure 3 and Table 3 respectively.

$$\text{Average percentage of BLEU method} = \frac{(73.325 + 30.325)}{(79.793 + 34.675)} * 100\%$$

$$\text{Average percentage of BLEU method} = \frac{103.65 * 100}{114.558} = 90.478\%$$

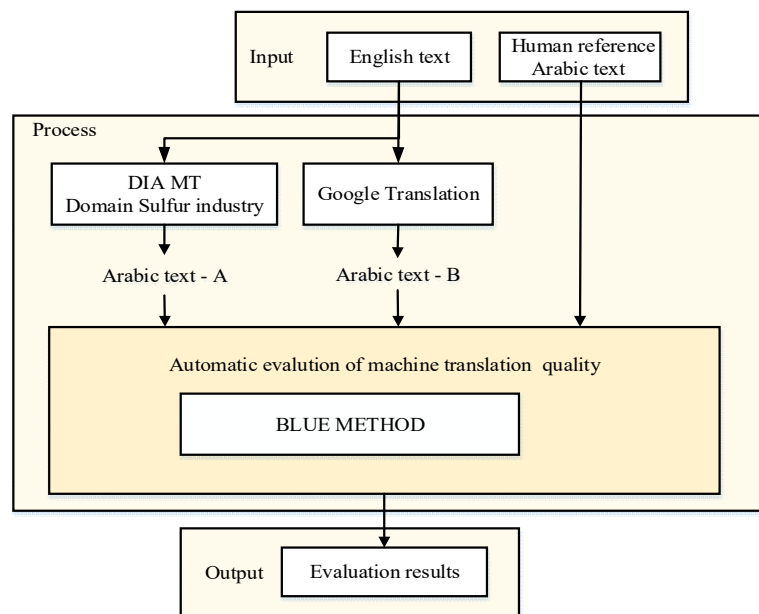


Figure 2. Block diagram of BLUE evaluation

Table 1. The scales for assigned fluency scores

Scales	Fluency
0.2	Incomprehensible
0.4	Non-fluent
0.6	Non-native
0.8	Good
1	Flawless

Table 2. Scales of scores used for assigned adequacy

Scales	Adequacy
0.2	None
0.4	Little
0.6	Much
0.8	Most
1	All

Figure 4 shows Summary of average precision, x-axis include terms of sulfur industry, phars less than six wordes, and sentences less then 50 wordes, the y-axis include the range of quality unit.

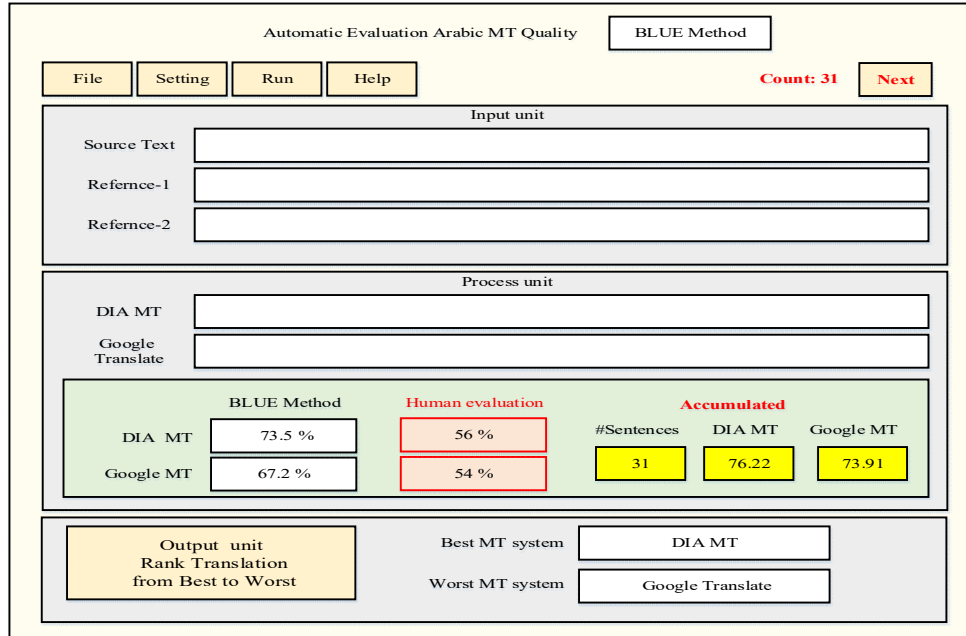


Figure 3. Results of MT systems evaluation

Table 3. Average precision for each type

Criteria	Translator	Terms	Phrase	Text	Average precision	Percentage method (%)
DIA MT	BLEU meth.	0.81	0.75	0.64	0.733	73.325
	Human tran.	0.85	0.80	0.73	0.793	79.793
Google	BLEU meth.	0.19	0.34	0.38	0.303	30.325
	Human tran.	0.24	0.39	0.41	0.347	34.675



Figure 4. Summary of average precision

#### 4. CONCLUSION

The recent research concludes that the automatic evaluation of MT of the efficiency of domain sulfur industry for DIA system using BLUE technique for determining the technique of assessment is closer to that human assessment. Furthermore, the recent research, refers that many experiments about the effectiveness concerning both internet systems of MT (i.e., Google translator and DIA translator) to translate 1,200 English symbols of sulfur, terms and texts within the competence of the sulfur industry into Arabic have been conducted. Most of the applied techniques to assess automatically the accuracy of the conversion of scheme of MT are relied on contrasting between the texts of both applicant and the reference.

The obtained findings refer that the normal acquiescence accuracy concerning the system of DIA translator is almost about 73.325% in comparison with that accuracy concerned with the Google system of MT of nearly 30.325% if the BLUE technique is used. The BLUE method efficiency is about (90.478%) as compared with the human expert evaluator.

## ACKNOWLEDGMENTS




This work cooperation of Computer Center/Company and Department of English Language, Tikrit University.

## REFERENCES




- [1] Y. K. Hussein, D. A. Hameed, L. I. Kalaf, B. Rahmatullah, and A. T. Al-Taani, "Automatic Evaluating Russian-Arabic Machine Translation Quality Using BLEU Method," *Revista AUS* 25, pp. 155–162, 2019.
- [2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [3] D. A. Hameed, T. A. Faisal, A. M. Alshaykha, G. T. Hasan, and H. A. Ali, "Automatic evaluating of Russian-Arabic machine translation quality using METEOR method," in *AIP Conference Proceedings*, 2022, p. 040036, doi: 10.1063/5.0067018.
- [4] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016, doi: 10.48550/arXiv.1609.08144.
- [5] Y. Qin, Q. Wen, and J. Wang, "Automatic evaluation of translation quality using expanded N-gram co-occurrence," in *2009 International Conference on Natural Language Processing and Knowledge Engineering*, Sep. 2009, pp. 1–5, doi: 10.1109/NLPKE.2009.5313751.
- [6] M. Yang *et al.*, "Extending BLEU Evaluation Method with Linguistic Weight," in *2008 The 9th International Conference for Young Computer Scientists*, Nov. 2008, pp. 1683–1688, doi: 10.1109/ICYCS.2008.362.
- [7] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the second international conference on Human Language Technology Research*, 2002, pp. 138–145.
- [8] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, "Evaluating English to Arabic Machine Translation Using BLEU," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 4, no. 1, pp. 66–73, 2013.
- [9] C.-Y. Lin and E. H. Hovy, "Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics," in *Proceedings of HLT-NAACL*, 2003, pp. 71–78.
- [10] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Comparative Study Between METEOR and B.L.E.U Methods of MT: Arabic into English Translation as a Case Study," (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, pp. 215–223, 2015.
- [11] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Evaluating Arabic to English Machine Translation," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 5, no. 11, pp. 68–73, 2014.
- [12] G. W. Blackwood, M. Ballesteros, and T. Ward, "Multilingual neural machine translation with task-specific attention," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3112–3122, doi: 10.48550/arXiv.1806.03280.
- [13] N. Adly and S. Al Ansary, "Evaluation of Arabic Machine Translation System Based on the Universal Networking Language," in *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems*, 2010, pp. 243–257.
- [14] T. Hailat, M. N. Al-Kabi, I. M. Alsmadi, and E. Al-Shawakfa, "Evaluating English to Arabic machine translators," in *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Dec. 2013, pp. 1–6, doi: 10.1109/AEECT.2013.6716439.
- [15] B. Babych and A. Hartley, "Extending the B.L.E.U MT evaluation method with frequency weightings," *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004.
- [16] A. A. Malik and A. Habib, "Qualitative Analysis of Contemporary Urdu Machine Translation Systems," *NLPAR@LPNMR*, pp. 27–36, 2013.
- [17] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of Bleu in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006, pp. 249–256.
- [18] A. Alqudsi, N. Omar, and K. Shaker, "A Hybrid Rules and Statistical Method for Arabic to English Machine Translation," in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, May 2019, pp. 1–7, doi: 10.1109/CAIS.2019.8769545.
- [19] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 13th IWSLT evaluation campaign," *IWSLT*, 2016.
- [20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019, doi: 10.48550/arXiv.1904.09675.
- [21] B. Babych and A. Hartley, "Meta-evaluation of comparability metrics using parallel corpora," *arXiv preprint arXiv:1404.3759*, 2014, doi: 10.48550/arXiv.1404.3759.
- [22] B. Babych *et al.*, "Training, Enhancing, Evaluating and Using MT Systems with Comparable Data," *Springer, Cham*, 2019, pp. 189–254.
- [23] R. Ananthkrishnan, P. Bhattacharyya, M. Sasikumar, and R. M. Shah, "Some issues in automatic evaluation of English-Hindi MT: More blues for bleu," in *Proceedings of the ICON*, 2007, pp. 1–8.
- [24] H. Azarbyoad, A. Shakery, and H. Faili, "A learning to rank approach for cross-language information retrieval exploiting multiple translation resources," *Natural Language Engineering*, vol. 25, no. 3, pp. 363–384, May 2019, doi: 10.1017/S1351324919000032.
- [25] A. Gupta, S. Venkatapathy, and R. Sangal, "METEOR-Hindi: Automatic MT evaluation metric for Hindi as a target language," in *Proceeding of the International Conference on Natural Language Processing Language*, 2010, pp. 1–10.
- [26] I. W. Kamal, H. A. Wahsheh, I. M. Alsmadi, and M. N. Al-Kabi, "Evaluating Web Accessibility Metrics for Jordanian Universities," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, pp. 113–122, 2016.

## BIOGRAPHIES OF AUTHORS






**Huda Mohammed Lateef**    she is lecturer in University of Fallujah. She has Master's degree in information technology from Universiti Utara Malaysia (UUM). She can be contacted at email: [hudamohammedlateef@uofallujah.edu.iq](mailto:hudamohammedlateef@uofallujah.edu.iq).






**Ahmed Mutar Awad**    he is an employee at the General Directorate of Education, Anbar Governorate, Ramadi, Iraq. Since 2006 graduated from the University of Anbar. Certificate from the University of Anbar, Iraq and a master's degree from Osmania University/Nizam College, Hyderabad, India, all in Computer Science and Information Technology. His research interests focus on computers and networks. He can be contacted at: [amaacs2@gmail.com](mailto:amaacs2@gmail.com).






**Diadeen Ali Hameed**    he is a lecturer in the Department of Electrical Engineering, Tikrit University, Tikrit, Iraq. He received the B.Sc. and Assistant Professor of Electrical Engineering at University of Tikrit, Iraq. He is an associate professor at the Department of Electrical Engineering, Al-Sherqat Engineering College, Tikrit University, Iraq, where he has been a faculty member since 2006. He graduated with a first-class honours B.Eng. degree at University of Mosul/Iraq and the M.Sc. degrees from University of Yurmouk, Jordin, all in Computer science and information technology. His research interests are in the area of computer and electrical engineering. He can be contacted at email: [diaa@tu.idu.iq](mailto:diaa@tu.idu.iq).



**Ghanim Thiab Hasan**    he is an Associate Professor at the Department of Electrical Engineering, Al-Sherqat Engineering College, Tikrit University, Iraq, where he has been a faculty member since 2006. He graduated with a first-class honours B.Eng. degree in Electrical and Electronic Engineering from Belgrade University, Serbia, in 1984, and M.Sc. in Electrical Engineering from Belgrade University, Serbia in 1986. His research interests are primarily in the area of electrical and electronic engineering. He can be contacted at email: [ganimdiab@yahoo.com](mailto:ganimdiab@yahoo.com).



**Tahseen Ameen Faisal**    he is an Assistant Professor at the Department of English language, College of Basic Education- Shirqat, Tikrit University, Iraq. He was a faculty member in the Department of Translation, College of Arts from 2007-2017. He graduated from the University of Mosul, Iraq in 1988, and M.A. in translation studies from the same university in 2000. His research interests are primarily in the area of linguistics and translation. He can be contacted at email: [Tahseen.faisal@tu.edu.iq](mailto:Tahseen.faisal@tu.edu.iq).