International Journal of Language and Literary Studies
Volume 5, Issue 2, 2023
Homepage : http://ijlls.org/index.php/ijlls

# Effect of Diacritics on Machine Translation Performance: A Case Study of Yemeni Literature

**Saleh Abduh Naji Ali Khoshafah**
*PhD research scholar, Dept. of English, Faculty of Languages, Sana'a University, Yemen*
salehkhoshafah2023@gmail.com
**Ibraheem N.A. Tagaddeen**
*Associate professor of Critical Theory and Translation, Dept. of Translation, Faculty of Languages, Sana'a University, Yemen*
ibtaj2007@gmail.com
*DOI*: https://doi.org/10.36892/ijlls.v5i2.1342

**APA Citation**: Khoshafah , S. A. N., & Tagaddeen , I. N. . (2023). Effect of Diacritics on Machine Translation Performance: A Case Study of Yemeni Literature . *International Journal of Language and Literary Studies*, *5*(2), 324–342. https://doi.org/10.36892/ijlls.v5i2.1342

***Received:***
*10/05/2023*

***Accepted:***
*09/07/2023*

***Keywords:***
*Diacritics, Machine translation (MT), Yemeni literature*

***Abstract***

*Many Arabic texts are written without diacritics, which can present difficulties for machine translation programs due to the high level of homography. Homographs are words that are spelled identically but have different meanings and are often pronounced differently. To avoid the problem of homography, words need to be diacriticized. The main objective of this study is to assess the ability of machine translation (MT) in rendering diacritical words from Arabic into English, with a focus on translating Yemeni literature into English. This study compares the translations of three MT programs: Reverso, Systran Translate, and Free Translation Online, to determine which program is closest to the original meaning of the source language texts. Additionally, the study aims to identify some causes behind errors in translating diacriticized words that result from these programs.*

*To achieve these aims, descriptive, analytical, and comparative methods were used by the researcher. Three common and modern MT programs - Reverso, Systran, and Free Translation Online - were selected to translate some diacriticized words. Excerpts with their contexts were taken from two Yemeni works: The Hostage (Ar-rahinah) (الرهينة) by the famous Yemeni writer Zayd Muttee Dammaj and the Yemeni book Yemeni Wealth from Popular Proverbs الثروة اليمنية من الأمثال الشعبية by the Yemeni writer Muhammad Al-Adimi. These samples were inserted into the MT programs for electronic translation and then analyzed and discussed qualitatively and quantitatively.*

*The study concluded that MT encountered problems with diacritics in Arabic texts; as a result, most of the time MT programs failed to recognize diacritics on letters. Thus, most of the program's translation results were incorrect and did not match the original meaning. It was also found that the Free Translation Online program produced the fewest errors of the three programs, while Systran mistranslated all of the diacriticized excerpts. These errors can be attributed to the absence of programs that contain the diacritic system of Arabic.*

## 1. INTRODUCTION

People in the same society communicate with each other by their mother tongue. However, those who are from different languages and cultures need an intermediator in order to communicate the meaning and translation is no exception. Therefore, translation played a crucial and important role in human beings life in various fields, for instance scientific, business, commercial, medical, legal and literary. At the age of modern technology, artificial intelligence came to make this easier than ever. The recent years have witnessed a lot of remarkable developments in many fields due to technological and scientific advancement. As a result, translation tools, methods and techniques were developed. The simple tools of translation (pen, dictionary, and notebook) have been replaced by electronic tools of translation. Recently the world witnessed the appearance of numerous MT programs especially for European languages on the market. Although these programs' quality is not generally good, the demand for these programs is very high. Moreover, the Internet has increased the need for MT which came to existence with the scientific advances in computational linguistics and invention of computers.

Thunes (2011) indicated that the first attempts of modern MT started in 1949 by Weaver who proposed using computers in translation for the first time. However, it was reported that automatic MT still lacked quality as compared to human translators. Therefore, the attempts continued to develop MT in the field of computational linguistics up to 1960. They wanted to create computational models that could match human performance. This generation of systems depended on an interlingual approach. Nevertheless, MT was criticized because of its slowness, lack of accuracy and cost compared to manual translation. By the late 1970s, the focus has been moved to transfer approach. Basically, the first steps of MT depended on the old method, direct approach of MT at word level.

## 2. Statement of the problem

MT has entered widely the life of modern person and became a part and parcel of his/her daily activities. This technology has been appreciated by numerous people, especially Internet users. Many MT issues have been discussed for example, MT approaches, MT problems, MT quality, etc. However, the study in hand will focus on the effect of diacritics on MT performance within the context of Yemeni literature translation. It aims to check MT ability to translate Arabic diacriticized words into English and highlight the causes that stand behind MT errors in translating such words.

## 3. Objectives of the study

The study aims to

- Assess the ability of MT in rendering diacritic words from Arabic into English with special reference to translating Yemeni literature.
- Investigate within a comparative framework the translations of the three MT programs *(Reverso, Systran Translate and Free Translation Online)* to find out which translation is close to the original meaning of the source texts.
- Identify the causes that stand behind errors of translating diacritics through MT programs (*Reverso*, *Systran Translate* and *Free Translation Online*).

## 4. Questions of the study

The study attempts to find answers to the following questions:

- Are MT programs able to render diacriticized words from Arabic into English adequately?
- Which one of the three mentioned MT programs can produce the closest translation to the source language meaning?
- What are the causes that stand behind errors of translating diacritics resulting from MT?

## 5. Significance of the study

As a matter of fact, diacritics play a very important role in understanding the lexical elements especially in Arabic language which has its own peculiarities. These features make it distinguished from other languages including English language. Thus, the role that diacritics play in adapting the difficulty of Arabic lexemes is an essential issue in this rich language. It is linguistically known that Modern Standard Arabic is not the only language variety used in the Arab countries. There are also many Arabic varieties like dialects, slangs, colloquial, etc. and Yemeni Arabic is an important part of these Arabic varieties. It can be said that diacritics and contexts should be taken into consideration by translators during the process of translating. Regarding MT, the problem is more serious due to the lack of a diacritic system in these programs. As a result, most of the time MT programs produce distorted and deviated translations and this is one of disadvantages of such programs. For this reason, this kind of study is insistently needed as it will be of great assistance to Arabic language translators. Not only that, but according to the best knowledge of the researchers, the research and studies done in this field of study are very rare. Due to all these issues, the researchers headed to perform this piece of study trying to cover some gaps which were not covered by other researchers and to come out with something new and valuable.

## 6. Delimitations of the study

This study is mainly focused on the diacritics translation problems from Arabic into English resulting from MT. Due to the limitation of time, some diacriticized examples were taken within their contexts from two works of Yemeni literature: *The Hostage" (Ar-rahinah الرهينة)* by Dammaj and *Yemeni Wealth from Popular Proverbs* (*الثروة اليمنية من الأمثال الشعبية*) by Al-Adimi. Besides, this study is confined to three common MT programs. In other words, the sample selected will be translated by using the programs *Reverso, Systran Translate* and *Free Translation Online*. In addition, this study will be carried out during the academic year 2022/2023.

## 7. Methodology of the study

The current study is descriptive analytical and comparative in nature. To answer the main question of the study: *Are MT programs able to render diacriticized words from Arabic into English adequately*? two types of samples were used: the first one is the MT programs which are *Reverso, Systran Translate* and *Free Translation Online*. These programs are modern and dependent on new MT approaches such as neural and hybrid MT. The second sample includes some examples excerpted from two Yemeni literary works which are *The Hostage" (Ar-rahinah الرهينة)* by Dammaj (1984), which represents one of the most prominent literary Yemeni works and *Yemeni Wealth from Popular Proverbs* (الثروة اليمنية من الأمثال الشعبية)

by Al-Adimi (n.d), which includes Yemeni popular proverbs. The sample chosen has focused on the homographic words which convey more than one meaning if not diacriticized. However, to avoid the problem of ambiguity, only diacriticized examples have been chosen as a sample and then were put into the three MT programs mentioned above for the purpose of translation online. Then, the programs' translations were copied, analyzed and discussed qualitatively and quantitatively in order to find out the errors made by these programs in translating diacriticized words, and the causes of such errors. After that, the results of MT programs were also compared with each other to identify the best program of the three in
translating diacriticized excerpts from Arabic into English.

## 8. Literature review

Theoretically speaking, translation from one language to another has invested in modern technology, taking help of computer advances. There are a lot of MT systems nowadays, each uses different MT approaches.

### 8.1 Machine translation (MT) approaches

There are different approaches of MT as presented in *figure 1* below.

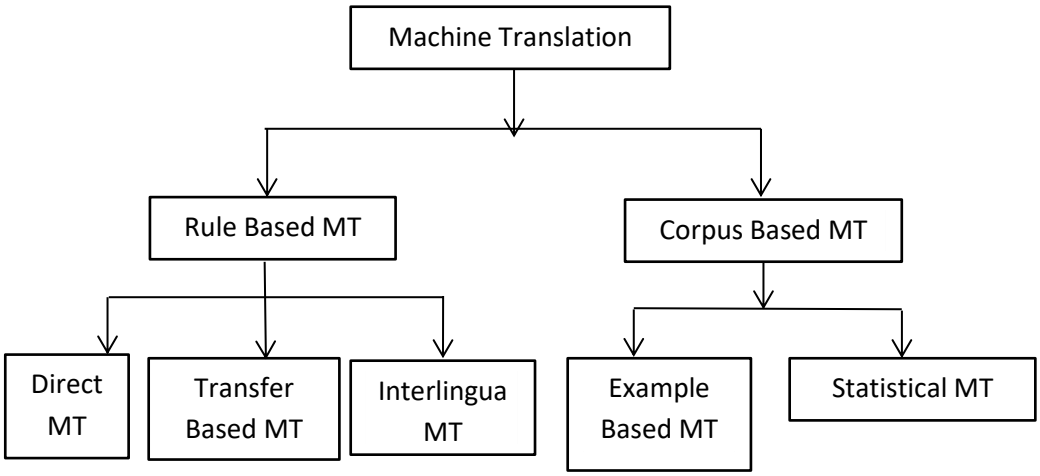


**Figure: 1** Different approach to machine translation (Kharb et al., 2017, p. 8)

#### 8.1.1 Direct approach

Direct approach is an old method in the history of MT. It uses bilingual dictionary to translate the source language text word-by-word. Yet this system lacks intermediate stages in translation processes and it has still a more primitive software design (Hutchins & Somers, 1992). According to Jurafsky and Martin (2000), each entry in the dictionary can be seen as a small program which is responsible for translating a single word. In addition, shallow morphological analysis and morphological generation can be applied.

#### 8.1.2 Transfer-based MT approach

This method analyzes the source sentence by parsing it to produce the syntactic representation of a source language sentence. The results are converted into an equivalent target language-oriented representation. It then generates the final text that is similar to the meaning of the original sentence (Okpor, 2014).

### 8.1.3 Interlingual MT approach

In this method, the source text is transformed into an interlingual language (language neutral) representation that is independent of any language. In other words, the source language text is transformed into a highly abstract representation that involves all the essential syntactic and semantic information that can then be converted into several target languages. The target language is then formed out of the interlingua. Thus, in multilingual MT systems, adding a new language would take less effort (Okpor, 2014).

### 8.1.4 Example-based MT

It requires parallel corpora. The already translated examples are used as knowledge to the system. This approach takes the information from the corpora for the three stages of translation; analysis, transfer and generation. Example-based MT systems take the source text and find the most similar examples from the source text in the corpora. The next step is to retrieve equivalent translations and create the final target edition (Sinhal, 2014).

### 8.1.5 Statistical MT approach

This approach looks up the source text in the parallel corpus to find its probability distribution function value with every sentence present in the target language for translation (Kharb et al., 2017). The statistic approach is based on probability of all the possible cases without any linguistic knowledge.

## 8.2 Diacritics in Arabic

Arabic language is based on alphabetical system (impure abjad) which contains 28 letters. It includes short consonants and long vowels which are represented by letters. It has three long vowels (أ a, و w, ي y). Arabic also uses diacritics *tashkil* or vowel marks. As for diacritics, they are phonetic aids and phonetic guides. It is worth mentioning that among the important characteristics of Arabic language today is the absence of diacritics. Diacritics are described in IPA (1999) as "small letter-shaped symbols or other marks which can be added to a vowel or consonant symbol to modify or refine its meaning in various ways" (p. 15). In addition, Hssini and Lazrek (2011) state that "a diacritic is a sign accompanying a group of letters or one letter, as the circumflex accent "^" on the "a" producing "â" (p. 1). Diacritics have different positions; some are placed above the letter; some can also be placed below the letter.

### 8.2.1 Types of diacritics in Arabic

The Arabic text is sometimes diacriticized by various short vowels. The main diacritics that are added to the Arabic letter are three short vowel sounds, /a, u, i/. For example, the letter (ب) has three main diacritics *harakat* "بَ /ba/, بُ /bu/ and بِ /bi/" above and below it (Abbad & Xiong, 2020). To begin with the first diacritic, *fatha* is a small diagonal line placed above a letter. It refers to the opening of the mouth and it represents the short vowel /a/ such as "بَ /ba/". The second diacritic is *dammah*, it is a small-curl like mark which is put above a letter to represent /u/, for example the *dammah* on the letter "بُ /bu/". *Kasrah* is put below a letter and it designates a short /i/ as in "بِ /bi/" (Lameris, 2021). Arabic language can also be characterized by other diacritics such as *sukūn*; it is a circle-shaped diacritic put above a letter. It indicates a silent letter for example, (سْ /s/). *Shaddah* or consonant gemination mark exists above the consonant letter which is to be doubled. It indicates consonant doubling or extra length, i.e. it

denotes stressing the letter. For example, in the word (مُدَرِّسة) which means 'female teacher', the pronunciation of the letter (رِّ /rr/) is doubled. Hence, it is different from the word (مَدَرَسة) which means 'school' in Arabic, in which the letter (رَ /r/) is diacriticized with *fatha*.

Moreover, Arabic has diacritical marks (*harakat*) which are called nunation (*tanwīn*); they occur on the final consonant of a word to indicate case for indefinite nouns, and they are pronounced as their respective vowel preceding an /n/. To be clear, *tanwīn* refers to final post-nasalized or long vowels in which the diacritics *fatha, dammah,* and *kasrah* are doubled at the end of the word to indicate (-*un*) in nominative case such as "ٌ /*tun*/", (-on) in accusative case such as "ً /ton/" and (-in) in genitive case as in (ٍ /tin/) (Lameris, 2021). The following table summarizes diacritics for the Arabic letter (س):

| Diacriticized letter (س) | Name of diacritic | Pronunciation |
|---|---|---|
| سَ | *Fatha* | /sa/ |
| سُ | *Dammah* | /su/ |
| سِ | *Kasrah* | /si/ |
| سًا | *Tanwin fatha* | /sun/ |
| سٌ | *Tanwin dammah* | /son/ |
| سٍ | *Tanwin kasrah* | /sin/ |
| سْ | *Sukun* | /s/ |
| سّ | *Shadda* | /ss/ |

**Table 1**: *The main Arabic diacritics for the Arabic letter (س)*

### 8.2.2 Function of diacritics in Arabic

Although they are omitted in most Arabic texts, diacritics perform a crucial role in disambiguating the meaning between two homographs. In some texts where the interpretation is important such as the Qurān, legal, or learners' texts, diacritics are usually placed on words; they differentiate slightly sounds. In other words, they are written in order to facilitate learning Arabic for foreigners and children (Neme & Paumier, 2020). The same glyph in writing Arabic can represent many letters and; without short vowels, the same word may represent multiple meanings. As a result, this causes reading difficulties because of confusion between consonants of the same shape (Hssini & Lazrek). When the context that surrounds homograph does not sufficiently disambiguate it, diacritics are required to be added to a word in a sentence. Readers' interpretations of undiacriticized Arabic in reading aloud depend primarily on the context and the sentence, but their accuracy improved when diacritics are present (Hermena et al., 2021).

Hallberg (2022) investigated the way in which Arabic diacritics are used. It used quantitative corpus linguistic methods to identify diacritization in a 72-million word corpus consisting of book of various genres. The number of diacritics used in children's literature and poetry vary considerably, while texts of normal prose include a narrow range of limited use of diacritics. A study was also conducted by Boudchiche and Mazroui (2015) to assess the level of ambiguity caused by the absence of Arabic diacritics in texts. This study followed a statistical method and was carried out based on four indicators: the root, the lemma, the stem and the part of speech tag of the word. A large diacriticized corpus was used; it included more than 80 million words collected from several sources. The study showed that diacritics are important for the meaning and their absence is the main cause of ambiguity. To resolve the problem of diacritics, Abo Bakr et al. (2008) suggested a statistical approach for diacritizing case-ending of an Arabic word using Support Vector Machines. Support Vector Machines gives the best results for many of natural language processing tasks, such as part of speech

tagging. This approach is automated and practical. The results can be useful in some applications that require real-time diacritization. The results of evaluating this system's performance showed that the technique is highly accurate with 95.3% accuracy and 82% F-measure.

### 8.2.3 Diacritics and translation

Diacritics are not only the issue of script, but also a combination with the spelling rules of Arabic. Arabic constitutes a challenging language due to its complex linguistic structure. Its script is mostly written without diacritics. Native speakers are able to disambiguate the intended meaning even if the text is not diacriticized. However, Fadel et al. (2019) found that absence of these diacritics or dropping case-ending sometimes pose problems for foreigners, children and some translators because diacritics are written purposefully to convey certain information about meaning based on its place within a sentence. Translating diacritics or short vowels can be evaluated at the character and word level. Some diacritics can change the word syntax for example to change a word from a part of speech into another, such as the homograph "حجر" can be a noun and a verb based on the vowels on it. If the word "حَجَرْ" is diacriticized with *fatha* on the characters (ج ,ح), it can be a noun to mean a "stone" whereas the word "حَجَّرَ" with *shaddah* (gemination) is a verb in past tense, which means "fossilized". Different diacritical marks also can change the word meaning, i.e., syntactic, or semantic analysis of one word may lead to several possible various word translations (Fadel et al.).

"The detection of case-ending diacritics is treated as a syntactic problem whereas detecting the internal diacritics is treated as a morphological problem" (Abo Bakr et al., 2008, p. 1). In translating Arabic texts, researchers attempt to design software that can analyze, understand and generate language, so that one will be able to address a computer as if addressing another person (Abusamrah, 2015). It can be said that the problem of diacritics is a very challenging one even to native speakers of Arabic due to the many subtle issues in determining the correct diacritic for each character because of lack of practice.

## 8.3 Homographs in Arabic

"Homographs are words that are written in the same way but are pronounced differently and have different meanings" (Palmer, 1984, p. 101). The word "close" is an example of homograph; it has different meanings according to its pronunciation, "close" /klous/ (adj.) and "close" / klouz/ (v). For Pyles (1971), homograph is a term used in semantic analysis to indicate lexemes that are written alike but may or may not be pronounced similarly and have different meanings. Hermena et al. (2021) also contributed and classified homographs into two main types; dominant representations and subordinate representations based on the process of diacritization.

The dominant representation is the most frequent meaning of a homograph, for example in Arabic the orthographic form "صوت" has the dominant meaning "voice or sound" because it is the familiar and frequent representation of the homograph "صوت". Many readers resorted to dominant meaning in case of absence of diacritics. However, it has a subordinate meaning of the same orthographic form which is "vote". The subordinate meaning can be instantiated by recognizing the diacriticized word "صَوَّت". It can be assumed that subordinate words are less frequent than the dominant representations. Thus, both representations, dominant and subordinate, are of the same orthographic form "صوت" /saut/ but they are phonologically and

semantically different. This proved that diacritics have an important role in altering the orthographic representation of the word. Hutchins and Somers (1992) confirmed that some homographs are more prevalent than others. These homographs can be disambiguated according to the text type. The unusual use of homographs is excluded from the dictionary unless it is appropriate to specific topic field of the texts to be translated. Homography and polysemy are treated alike by MT.

### 8.4 Machine translation (MT) systems

***Microsoft Translator*** is a multilingual MT cloud service developed by Microsoft. *Microsoft Translator* has the ability to offer text and speech translation. It is a part of Microsoft Cognitive Services. This MT program was developed and provided with modern approaches such as statistical methods. By this method, algorithms are trained to understand translated similar texts. In 2016, *Microsoft Translator* was updated to offer deep neural MT as a method in translation in speech languages. It adopted neural MT because this approach provides better translations than statistical MT approach. In addition, this translation program uses transliteration and bilingual dictionary to look up words to provide alternative translations and display them in examples. *Microsoft Translator* also can support several speech translations ("Microsoft Translator," 2023).

***Reverso*** is a company in language tools such as translation tools and language services. These aids include online MT which is based on neural MT, contextual dictionaries, spell and grammar checking and conjugation tools. *Reverso* has been used since 1998 and it has over 96 million users. It can also edit and review translated content through a simple interface. It supports several languages including Arabic, Chinese, English, Hebrew, French, Italian, Spanish, Ukrainian and Russian. In 2013 *Reverso Context* was released; it is a bilingual dictionary tool based on a big corpus and machine learning algorithms ("Reverso (language tools)," 2022).

***SYSTRAN*** is a fast and multilingual system which was developed by the European Commission. It is founded by Dr. Peter Toma in 1968 and it is one of the oldest MT companies. Anastasiou (2011) provided a summary of *SYSTRAN* program in the following lines:

> SYSTRAN debuted in the market in the 1970s as a robust RBMT system and has been a successful commercial tool still to the present days. Currently SYSTRAN is a hybrid system, which is claimed to combine the predictability and consistency of RBMT systems with the fluency of SMT systems (Senellart, 2009). Another strong claim by the company for commercial targets but also of interesting academic value is that SYSTRAN possesses a learning module, which is used for system training. It extracts sentences from corpora, but rules may get adapted with repetitive use to fit translation domains, so people or parties using the system should gain speed and accuracy of translation with the long-term use of SYSTRAN. (p. 122)

*SYSTRAN* was developed to have the capability of adding multiple meaning resolutions to the system at different levels by means of semantic categorization. This enables SYSTRAN to specially disambiguate words in the case of multilingual translations (Toma, 1977).

## 9. Data analysis

Yemeni literature includes modern standard Arabic and dialects. Some of these works include words which convey multiple meanings due to diacritics. To assess MT performance

in realizing diacritics on Yemeni Arabic words and identifying the real meaning of word, some examples with diacriticized words from Yemeni works will be discussed below:

*Example*: *1*

| Source Text | | سُلِّمت لي عدة حزم من القات (Dammaj, 1984, p. 80) |
|---|---|---|
| Machine translations | *Reverso* | I was handed several bundles of khat |
| | *Systran Translate* | She handed me several packets of khat |
| | *Free Translation* | She gave me several packets of Qat |

The Arabic verb "سلمت" is heterophonic-homographic verb that has different pronunciations in active and passive, i.e. "سَلَّمَت", "she submitted" or "she gave" as an active form whereas the verb "سُلِّمَت" is a passive form and it means " it was given" or "it was handed". The underlined and diacriticized Arabic verb is in the passive form. In the Arabic system, the passive form of a verb is diacriticized with *dammah* on the first letter. The sentence context that surrounds the homograph "سلمت" does not sufficiently disambiguate it, thus diacritics are essential for the meaning of the verb "سلمت". Practically speaking, the homograph "سُلِّمَت" was input into the three programs (*Reverso, Syatran* and *Free Translation Online*) and only *Reverso* system translated it correctly; it realized the function of diacritics so it translated the word "سُلِّمَت" as passive voice. On the contrary, the two programs *Systran* and *Free Translation Online* failed in translating diacritics; they transformed the source passive verb into active form though it is diacriticized.

*Example: 2*

| Source Text | | أَلِف ولا تقطع (Al-Adimi, n.d, p. 21) |
|---|---|---|
| Machine translations | *Reverso* | Alf and don't cut |
| | *Systran Translate* | A Thousand and No Interruptions |
| | *Free Translation* | A thousand and do not cut |

As seen in example 2, the first Arabic word is an imperative verb. Without diacritics, the word "ألف" is so ambiguous and it may have multiple meanings since it can be provided with various short vowels. If it is vowelized as "أَلِفْ", it means the first letter in Arabic (أ), "Alf", but the same word "أَلِّفْ" with *Shaddah, gemination mark on the letter* (ل) indicates another meaning. According to context here, the word "أَلِّفْ" means "make a habit with someone". Furthermore, "أَلْفْ" means "thousand". Nevertheless, the three MT systems interpreted diacritics incorrectly; *Reverso* transcribed the dialectal verb "أَلِّفْ". However, *Systran* and *Free Translation Online* translated it wrongly into "a thousand". They faced difficulty in translating dialectal words. The suggested translation for the whole sentence "أَلِّف ولا تقطع" is "make a habit with someone but don't stop it" or "if you make a habit with someone, don't stop it".

*Example: 3*

| Source Text | | المُوْدِّع نُص رجَال ( Al-Adimi, n.d, p. 61) |
|---|---|---|
| | *Reverso* | Deposit Text Men |

| | | |
|---|---|---|
| Machine translations | *Systran Translate* | The depositor is half men |
| | *Free Translation* | The depositor is half a man |

The two words "الـمُـوْدِّع" and "نُـص" in the Arabic edition are also examples of homographs. *Reverso* program translated the whole sentence literally. In other words, *Reverso* was not able to recognize the correct diacritics; this program translated it as a word without the diacritic *shaddah* on the letter (د) "المُوْدِع". As a result, "المُوْدِّع" and "نُص" are heterophonic-homographs. Each has its own meaning; "المُوْدِّع" means "the person who depends on others to do his duties" whereas "الـمُـوْدِع" with the short vowel *kasra* under the letter (د) means "depositor". In addition, *Reverso* mistranslated the word "نُص" into "text". The word "text" is an equivalent for the word "نَص" with the diacritic *fatha* on the letter (ن); however, the source word "نُـص" with *dammah* on the letter (نُ) can be translated into "half". These three MT programs have failed to arrive to the close meaning of the original text and they distorted the translation of the word "المُوْدِّع" in the mentioned context as financial and banking kind of translation (depositor). On the other hand, the Yemeni cultural meaning for the whole sentence is that "the person who depends on others to do things for him is a half man". The connotative meaning for this Yemeni proverb is that if you depend on others, they will not do things better for you as you do for yourself or the person should help himself and do not wait others to help him. The best comparable expression for this proverb in English is Bonaparte's quote "If you want a thing done well, do it yourself".

*Example 4:*

| | | |
|---|---|---|
| Source Text | | اليوم بأخيك, بُكرة فيك (Al-Adimi, n.d, p. 24) |
| Machine translations | *Reverso* | Today with your brother, early in you. |
| | *Systran Translate* | Today with your brother, early in you. |
| | *Free Translation* | Today with your brother, tomorrow you will |

The underlined word "بُكرة" is diacriticized with *dammah* on the letter (بُ) and this word is used more often in many Arabic dialects in many Arab countries like Yemen, Egypt, Iraq, Sudan, etc. The two programs *Reverso* and *Systran* failed completely in translating such word into English and this again ensures the assumption of this study which states that MT programs fail in translating diacriticized words especially Arabic dialects. On the other hand, *Free Translation Online* program succeeded in translating the word "بُكرة" correctly. This means that the translation of MT programs that have many errors at present time can be improved by feeding them with diacriticized words with their correct equivalents.

*Example: 5*

| | | |
|---|---|---|
| Source Text | | فوجدتُ صاحبي قد نام أو أنه تَصَنَّع ذلك (Dammaj, 1984, p. 114) |
| Machine translations | *Reverso* | I found my friend sleeping or he made it. |
| | *Systran Translate* | So I found my friend had fallen asleep or had made it |
| | *Free Translation* | And I found my friend had fallen asleep or that he had made it up |

The three translations of the programs are similar especially in translating the vowelized word "تَصَـنّع", but they translated it incorrectly into "made". There is no semantic relation between the source diacriticized verb "تَصَـنّع" and the programs' results "made". The word "تَصَنَّع" with *shaddah*, *gemination* on the letter (نّ) means "pretend". The three above-mentioned

MT systems faced the problem of the diacritic *shaddah* placed above the letter (نّ) in the word "تَصنَّع"; they could not recognize the real meaning of the word "تَصنَّع" so they translated it into "made" "صنَعَ".

*Example: 6*

| Source Text | | آلة ذات إطارات أربعة تُقِلُّ أكثر من شخص أو شخصين (Dammaj, 1984, p. 85) |
|---|---|---|
| Machine translations | *Reverso* | Four-frame machine transports more than one or two people |
| | *Systran Translate* | A four-frame machine that kills more than one or two people |
| | *Free Translation* | A four-framed machine carrying more than one or two people |

The tri-verb "تُقِلُّ" has a specific meaning due to the short vowels placed on it. The first letter "تُ" is vowelized with *dammah* and there is *dammah* and *shaddah* on the letter (لُّ). If the same orthographic word "تَقِلُ" has *fatha* on the first letter (تَ) and *kasra* on the second letter (قِ), the meaning will be different. Both *Reverso* and *Free Translation Online* were successful in translating the verb "تُقِلُّ" whereas *Systran* program's translation was very far from the source meaning "تُقِلُّ"; it added the letter (ت) for the verb "تقل" to become"تقتل", "kill".

*Example: 7*

| Source Text | | سَمِّن كلبك يأكلك (Al-Adimi, n.d, p. 242) |
|---|---|---|
| Machine translations | *Reverso* | Fatten your dog eating you |
| | *Systran Translate* | Your dog's margarine will eat you |
| | *Free Translation* | Your dog's margarine eats you |

The source sentence above is an imperative sentence which begins with the verb "سَمِّن". This verb has *shaddah* and *kasra* on the letter (مِّ). Based on the sentence context, the word "سَـمِّن" means "feeding the dog to become fat". The same word "ســمن" can be interpreted differently if it diacriticized as "سَـمْن"; it is a noun meaning "cooking fat" or "butter". It has been found that only *Reverso* program could realize the function of *shaddah* on the letter (مّ) in the word "سَمِّن" whereas *Systran* and *Free Translation* programs mistranslated the diacritics of the word "سَمِّن ". They translated it as noun "margarine".

*Example: 8*

| Source Text | | الجمل من جَمَّالِه (Al-Adimi, n.d, p. 42) |
|---|---|---|
| Machine translations | *Reverso* | Camels of its beauty |
| | *Systran Translate* | The Camel's Beauty |
| | *Free Translation* | The Camel's Beauty |

Homographic words occur when the letters appear without diacritics. For example, the word "جماله" is ambiguous and can be interpreted by different ways. When it is vowelized as "جَمَالِه", it implies a certain meaning "his/her beauty". However, if the word "جَمَّالِه" is provided with *shaddah* on the letters (مَّ), it indicates a dialectal word which can be translated into "camel's owner". Furthermore, the word "جِمَالَه" with *kasra* under the letter (جِ) and *fatha* on the letter (مَ) can mean "his camels". Thus, short vowels (diacritics) in Arabic disambiguated the meaning of the word "جـمــالــه". It is remarkable above that all the three MT programs mistranslated the word "جَمَّالّه" though it is diacriticized. They translated it as a standard Arabic

word "جَمَالِه", "beauty". For MT programs, the most frequent and dominant representation of the homograph "جماله" is "جَمَالَه", "beauty".

*Example: 9*

| Source Text | | (Dammaj, 1984, p. 44) وديوان النائب دائما مكتظ بالسُمّار |
|---|---|---|
| Machine translations | *Reverso* | The Deputy's Office is always overcrowded. |
| | *Systran Translate* | The Deputy's Office is always full of booths. |
| | *Free Translation* | And the office of the deputy is always crowded with nails. |

There is a large class of words that fall under the homograph category. The previous excerpt includes one of the homographic words; it is the word "السُـمّار". In Arabic, the word "السُمّار" which has *dammah* on the letter (سُ) and *shaddah* on the letter (مّ) is a plural noun (its singular is سَـامِر) and it means in English "talkative" or "the people who stay together talking and sometimes chewing Qat in a room for a long time at night". However, the word "السَمّار" is a kind of plant ("*Almaany Dictionary*," 2023). Regarding MT results, *Reverso* program did not identify the Yemeni cultural word "السُـمّار" in this context; consequently, *Reverso* deleted the word "السُـمّار" from its translation. Again *Systran* program translated the word "السُـمّار" into "booths" which is not related to the original meaning at all. Similarly, *Free Translation Online* made irrelevant translation and translated the word "السُمّار" into "nails".

*Example: 10*

| Source Text | | (Dammaj, 1984, p. 122) أَركُبتُ فوق حصان مقوس الظهر |
|---|---|---|
| Machine translations | *Reverso* | I got on top of an arched horse |
| | *Systran Translate* | I rode on a curved back horse |
| | *Free Translation* | I rode on a horse with an arch-back |

The verb "أُركُبتُ" in the source sentence is vowelized with *dammahs* on the first letter (أُ) and (كُ). Due to the diacritics placed on the verb "أُركُبتُ", it is in the passive form; the agent is unknown. The three programs transformed the passive verb "أُركُبتُ" into active form. Their translations are suitable only for the verb "رَكِبتُ". Thus, they distorted the meaning when they regarded the verb "أُركُبتُ" as an active form and translated it as "got on" and "rode". No one of the above-mentioned programs could recognize the function of the short vowels on the letters of the word "أُركُبتُ". This result indicates that diacritics may not be available in MT programs databases.

*Example: 11*

| Source Text | | (Al-Adimi, n.d, p. 341) لا حذر من قَدَرْ |
|---|---|---|
| Machine translations | *Reverso* | Don't be careful who you can. |
| | *Systran Translate* | La Hazer Min Values. |
| | *Free Translation* | Don't beware of fate. |

The options "fate", "destiny", "pot", "amount", "assume", "appreciated", "estimate" and "could" can be accepted for the Arabic word "قـدر" if it is out of the context and without diacritics. This means that the element "قدر" is a homographic word. It is clear from the source Arabic text that the underlined word "قَدَرْ" is a noun (complement) because it comes after the preposition "مِن". Prepositions in Arabic govern their complements to appear in the genitive

case. What determines the meaning of the homographic word is diacritics and context. The source word "قَدَرْ" is vowelized with *fatha* on the first two letters (د,ق) and *sukūn*. Accordingly, it can be translated as a noun "fate" or as a verb "could, was able to"; however, based on the context it has only one meaning which is "fate" or "destiny". *Reverso* and *Systran Translate* programs faced a difficult task to find the equivalent word in the target language to produce a clear message. *Reverso* translated the word "قَدَرْ" into the verb "can" which is not appropriate to the Arabic context. *Systran* also provided a poor translation when it translated the word "قَدَرْ" into "values". In regard to the third MT program *Free Translation Online*, it was different from the other two programs; its translation was correct because it offered the right equivalent for the word "قَدَرْ", "fate".

*Example: 12*

| Source Text | | لا تربط حمارك جنب حمار المُدْبِر (Al-Adimi, n.d, p. 342) |
|---|---|---|
| Machine translations | *Reverso* | Do not associate your ass with the ass of the mastermind |
| | *Systran Translate* | Don't tie your donkey next to the head donkey |
| | *Free Translation* | Do not tie your donkey next to the donkey of bad luck |

Example 12 shows one of the most popular Yemeni proverbs. This proverb includes the homographic word "المـدبر". To be precise, this word conveys many meanings such as "المُدَبِّر", "المُدْبِر", "المِدْبِر". The diacriticized word "المُدَبِّر" is an adjective and it means "planned" or "mastermind" whereas the two words "المُـدْبِر" and "المِـدْبِر" are dialectal words which are spoken in Yemen and can be translated into "hopeless" or "unlucky" person. It is worth mentioning that the word "المُدْبِر" also belongs to standard Arabic; it means "goer", i.e. it is the opposite of "comer". The vowelized word in the source text above "المُـدْبِر" was translated incorrectly by *Reverso* and *Systran* programs; however, *Free Translation* program's translation was close to the original text. This program could relatively identify the real function of diacritics added to the word "المُدْبِر".

## 10. Findings and conclusion

This study was conducted to assess MT efficiency in translating vowelized words from Arabic into English, so some examples were chosen purposefully; some examples are from standard Arabic and others from Yemeni dialects. The selected examples are provided with diacritics. After analyzing *Reverso*, *Systran* and *Free Translation*'s results, it can be said that translating diacritics from Arabic into English is considered as crucial and noticeable translation problem that can attract strongly the attention of the trans-editor and post-editor translators. Generally speaking, it has been found that the three MT programs, namely, (*Reverso, Systran Translate* and *Free Translation Online*) faced difficulties in translating Arabic diacriticized words into English as they committed serious errors in translating such words, though these words were inserted into these MT programs along with their contexts. It can also be shown that MT programs provided the most frequent translations (dominant meaning) for the given homographs and they ignored the subordinate representations of the homographs though the ambiguity was avoided by diacritics. For example, the word "سَمِّن" in the sentence "سـمِّـن كـلـبـك يـأكـلـك" was translated into its dominant and common meaning "margarine". This means that MT ignored the function of the diacritic *shaddah* which

determines the real meaning of the word "سَمّن". Furthermore, the process of analysis displayed that the three MT programs' qualities are different as shown in the following table:

| Number of examples | *Reverso* | | *Systran Translate* | | *Free Translation Online* | |
|---|---|---|---|---|---|---|
| | Errors | Percentage | Errors | percentage | Errors | percentage |
| 12 | 9 | 75 % | 12 | 100 % | 8 | 66.6 % |

**Table***: 2 MT programs' errors and percentages of translating diacriticized words*

As it is clear in table (2) above, *Systran* program mistranslated all the diacriticized words. *Free Translation Online* program is the best of the three programs though it could not translate eight diacriticized words; the most problematic areas for *Reverso* and *Free Translation Online* in this sample are dialectal words such as, "جَمَّـالـه" and " المُوْدِّع". Since modern Standard Arabic does not have orthographic representation of short letters and this poses problems for MT, it is necessary to determine homographic words phonologically or semantically and disambiguate word sense by means of diacritics. Errors of MT programs in translating diacritics are attributed to many causes; the most important of which is that these programs do not contain the diacritic system of Arabic in their memories, hence these programs cannot deal with diacritics properly. These programs are also used rarely in translating dialects and in translating diacriticized words so they cannot recognize diacritical marks easily. Finally, this study recommends that these MT programs and others should be fed with Arabic diacritics system of the words and this will consequently enable such programs to overcome the problems of translating Arabic diacritical words into English. Not only that, but it is also necessary to provide MT programs with diacritic systems of Arabic dialects in order to make MT programs identify the correct meaning of any inserted Arabic words especially homographs and come out with correct and appropriate translations.

## REFERENCES

Abbad, H., & Xiong, S. (2020). Multi-components system for automatic Arabic diacritization. In *Proceedings of the European Conference on Information Retrieval*, 341–355. https://doi.org/10.1007/978-3-030-45439-5_23

Abdulaal, M. A. (2022). Tracing machine and human translation errors in some literary texts with some implications for EFL translators. *Journal of Language and Linguistic Studies*, *18*(1), 176-191. https://files.eric.ed.gov/fulltext/EJ1328441.pdf

Abo-Bakr, H., Shaalan. K. & Ziedan, I. (2008). A statistical method for adding case diacritics for Arabic text. *Language Engineering Conference*. Ain Shams University, 225–234. www.tinyurl.com/muwfheq

Abu-Rabia, S. (2002). Reading in a root-based-morphology language: The case of Arabic. *Journal of Research in Reading*, *25*(3), 299-309.

Abusamrah, I. (2015). *Al-Farāhīdī Arabic diacrizer system* [Unpublished master's thesis]. Al- Quds University. https://core.ac.uk/download/pdf/287333184.pdf

Al-Adimi, M. (1989). *Yemeni wealth from popular proverbs* (2nd ed.). Beirut: Alsabagh Corporation.

Al-Iriyani, M. (1996). *Al-Mu'jam Al-Yemeni fi lughat wa-turath* (1st ed.). Damascus: Dar Alfekr.

*Almaany Dictionary*. (n.d.). Retrieved June 15, 2023, from https://www.almaany.com/ar/dict/ar-ar/السمار

Al-Salman, S. (2022). The effectiveness of machine translation. *International Journal of Arabic-English Studies,5*, 145-160. file:///C:/Users/window/Downloads/8-4.pdf

Al-Zebary, Y. T. (2012). *Lexical and structural ambiguity in machine translation* [Unpublished master's thesis]. University of Science & Technology. file:///C:/Users/window/Downloads/Lexical_and_Structural_Ambiguity_in_M ach-1.pdf

Anastasiou, D. (2011). Comparison of Systran and Google Translate for English-Portuguese. Retrieved June 1, 2023, from http://revistes.uab.cat/Tradumatica/article/248906

Baker, M. (1992). *In other words*. London: Routledge.

Baker, M. & Saldanha, G. (2009). *Routledge encyclopedia of translation studies.* London & New York: Routledge.

Ball, M. J. (2001). On the status of diacritics. *Journal of the International Phonetic Association*, *31*(2), 259–264. https://doi:10.1017/S0025100301002067

Boudchiche, M., & Mazroui, A. (2015). Evaluation of the ambiguity caused by the absence of diacritical marks in Arabic texts: Statistical study. In *International Conference on Information & Communication Technology and Accessibility*. https://doi: 10.1109/ICTA.2015.7426904

Carl, M., & Way, A. (Eds.). (2004). Recent advances in example-based machine translation. *Text, Speech and Language Technology*, *21*(4), Dordrecht: Kluwer Academic.

Catford, J. (1965). *A linguistic theory of translation*. London: Oxford University Press.

Dammaj, Z. M. (1984). *Al-Rahina* (The hostage) (1st ed.). Beirut: Dar Al adaab.

Dammaj, Z. M. (1994). *The Hostage* (M. Jayyusi & C. Tingley,Trans.). New York: Interlink Books.

Fadel, A., Tuffaha, I., Al-Jawarne, B., & Al-Ayyoub, M. (2019). Neural Arabic text diacritization: State of the art results and a novel approach for machine translation. *Proceedings of the 6th Workshop on Asian Translation*, 215-225. https://doi.org/10.18653/v1/D19-5229

Ghazala, H. (2015). *Translating culture: A textbook*. Jedda, Saudi Arabia: Konooz Al-Marifa.

Hallberg, A. (2022). Principles of variation in the use of diacritics (*taškīl*) in Arabic books. *Language Sciences, 93*, 1–15. http://www.doi.org/10.1016/j.langsci.2022.101482

Hermena, E., Bouamama, S., Liversedge, S., & Drieghe, D. (2021). Does diacritics-based lexical disambiguation modulate word frequency, length, and predictability effects? An eye movements investigation of processing Arabic diacritics. *PLoS ONE, 16*(11). https://doi.org/10.1371/journal.pone.0259987

Hssini, M. & Lazrek, A. (2011). Design of Arabic diacritical marks. *International Journal of Computer Science*, *8*(3), 262–271. file:///C:/Users/window/Downloads/Design_of_Arabic_Diacritical_Marks-1.pdf

Hutchins, W. J. (1986). *Machine translation: past, present, future*. Chichester (UK): Ellis Horwood.

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*. London: Academic Press.

Hutchins, W. J. (2010). Machine translation: A concise history. *Journal of Translation Studies, 13*(1-2), 29-70. https://cup.cuhk.edu.hk/chinesepress/journal/JTS13.1-2/JTS13.1-2_29-70.pdf

IPA (1999). *Handbook of the international phonetic association*. Cambridge: Cambridge University Press.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognitions* (2nd ed.). New Jersey: Prentice-Hall.

Kharb S., Kumar, H., & Chaturvedi, A. (2017). Efficiency of a machine translation system. In *International Conference of Electronics, Communication and Aerospace Technology*, 140–148. https://doi: 10.1109/ICECA.2017.8203660

Koehn, P. (2010). *Statistical machine translation*. Cambridge: Cambridge University Press.

Lameris, H. (2021). Homograph disambiguation and diacritization for Arabic text-to-speech using neural networks [Unpublished master's thesis]. Uppsala universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-446509

Microsoft Translator. (2023, May 15). In *Wikipedia*. Retrieved June 3, 2023, from https://en.wikipedia.org/wiki/Microsoft_Translator

Neme, A., & Paumier, S. (2020). Restoring Arabic vowels through omission-tolerant dictionary lookup. *Language Resources and Evaluation*, *54*(2), 487-551. file:///C:/Users/window/Downloads/Restoring_Arabic_vowels_through_omission-6.pdf

Newmark, P. (1988*). A Textbook of translation*. New York: Prentice Hall.

Okpor, M. D. (2014). Machine translation approaches: Issues and challenges. *International Journal of Computer Science Issues* (IJCSI), *11*(5), 159-165. https://www.ijcsi.org/papers/IJCSI-11-5-2-159-165.pdf

Palmer, F. R. (1984). *Semantics*. Cambridge: Cambridge University Press.

Pyles, T. (1971). *The Origin and development of the English language.* New York: Harcourt Brace Jovanovich, Inc.

Quah, C. (2006). *Translation and technology*. New York: Palgrave Macmillan.

Reverso (language tools). (2022, November 22). In *Wikipedia.* Retrieved June 2, 2023, from https://en.wikipedia.org/wiki/Reverso_(language_tools)

Sahin, Ö. (2015). *Consistency in the evaluation methods of machine translation quality* [Unpublished master's thesis]. Hacettepe University Graduate School of Social Sciences, Ankara.

Sekhri, O. (2019). Machine translation and the problems of translating cultural terms in the Arab World. *AL-MUTARĞIM Journal*, *19*(1), 239-261. file:///C:/Users/window/Downloads/OuidedSEKHRI-10.pdf

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.

Sinhal, R., & Gupta, K. (2014). Machine translation approaches and design aspects. *IOSR Journal of Computer Engineering*, *16*(1), 22-25. file:///C:/Users/window/Downloads/Machine_Translation_Approaches_and_Design_Aspects.pdf

Thunes, M. (2011). *Complexity in translation: An English-Norwegian study of two text types* [Unpublished doctoral dissertation]. University of Bergen, Bergen: Norway.

Toma, P. (1977). SYSTRAN as a multilingual machine translation system. *Overcoming the language barrier*, *1*, 569-581. https://aclanthology.org/www.mt-archive.info/70/CEC-1977-Toma.pdf

Venuti, L. (2004). *The translation studies reader*. London & New York: Routledge.

Vilar, D., Xu, J., D'Haro, L., & Ney, H. (2006). Error analysis of machine translation output. *International Conference on Language Resources and Evaluation*. Genoa, Italy, 697–702. http://www.lrecconf.org/proceedings/lrec2006/pdf/413_pdf.pdf

Watson, J. C. (1993). *A syntax of San*ʕ*ani Arabic*. Wiesbaden: Otto Harrassowitz.