A Tool for Measuring Linguistic Variations in Machine Translation: A Corpus Based Study

Maram Elsaadany

Institute of Applied Linguistics & Translation, Faculty of Arts, Alexandria University, Alexandria, Egypt

Maram.elsaadany@gmail.com

Sameh Alansary

Bibliotheca Alexandrina, Alexandria, Egypt
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
sameh.alansary@bibalex.org

Abstract — ***This paper is primarily a translation analysis of the Arabic and English morpho-syntactic structures using Biber's model (1988) and Stanford computer program. It is a corpus based quantitative study that has used 66 features out of the 67 that has been identified by Biber and the paper is in line with Biber's model and statistical procedure. The corpus selected for this thesis is Alice Monro's collection of short stories The Power of Love (1985) and its translation into Arabic, Masiret El Hob (2015) by Mohamed Tantawi. All the English and Arabic features are counted by the aid of a computer program, Stanford (2015). Stanford is a program that is used for annotating the chosen corpus and it is working on the morpho-syntactic levels of English and Arabic. The 66 features are classified into four factors for the English language and five factors for the Arabic. Only twelve features are counted manually in the Arabic analysis by the researcher herself. Finally, the findings reflect some differences between the two languages.***

Key words: ***Translation, Computational, Linguistic variations***

1 INTRODUCTION

Corpora are collections, usually electronic ones today, of texts. A 'parallel corpus' is a bilingual or multilingual corpus that contains one set of texts in two or more languages" [1]. There are altogether three types of parallel corpora, and their functions are different according to their different construction. The first type is the normal parallel corpus. This type contains only texts of language, usually source language, and their translation into another language, target language. The corpus of this study belongs to this type. The second type is the reciprocal parallel corpus. It contains not only the source texts in language A and their translation in language B, but also source texts in language B and their translation in language A. The third type contains only translations in different target languages. This type may be bilingual or multilingual.

Corpus-based translation studies are interested in how equivalence might be achieved and what kind of equivalence can be achieved, and in what context [2] Translation unit studies are interested in the alignment of translation units and their equivalents in a given parallel corpora, and how these equivalents can be re-used by other translators in the future translations, especially by those translators who have to translate into a non-native language where their intuition is often insufficient. The unit of annotations and the choice of the annotation scheme are crucial for the quality of this research. [3] has expressed that translation units are the smallest unit in translation and they are very useful for bilingual lexicography. [4] has stated that "parallel corpora are repositories of the translation units and their equivalents". Computational linguistics (CL) combines resources from linguistics and computer science to discover how human language works. Computational linguistics is a vital field in the information age. According to [5], computational linguists create tools for important practical tasks such as machine translation, speech recognition, speech synthesis, information extraction from text, grammar checking, text mining and more. [6] has stressed the idea that contrastive Analysis (CA) is a method that is connected to Contrastive Linguistics, which is considered a branch of linguistics that focuses on illustrating the differences and similarities among two or more languages at different linguistic levels as semantics, syntax, and phonology, [7]. [8] defines contrastive analysis as "the study of foreign language learning, the identification of points of structural similarity and difference between two languages".

The use of computerized text corpora and computer programs for the automatic identification of linguistic features made it possible to fulfill a study of this scope. The words in any text are all marked, or 'tagged', for their grammatical category, to facilitate automatic syntactic analysis. There are two main steps associated with automatic identification of the linguistic features. The first is to tag the grammatical category of each word, as a noun, verb, adjective, preposition, WH pronoun, etc. [9] has explained that this step requires a computerized dictionary so that the program can search for words in the dictionary and find their grammatical category. The tags resulting from this procedure provide the basis for the second step, which is identifying particular sequences of words as instances of a linguistic feature. For example, if a noun is followed by a "WH pronoun" and not preceded by the verb "tell or say", it can be identified as a relative clause; the sequence tell/say + noun phrase + WH pronoun might be either a relative clause or a WH clause. Working on the programs which can be used for the frequency counts of the features has spread over the years (1983- 1986).

Earlier Programs have been criticized by the lack of a dictionary; to identify linguistic features, they relied on small lists of words that were built into the program structure itself. These lists included prepositions, conjuncts, pronominal forms, auxiliary forms. Since these word lists were relatively restricted, the grammatical category of many words in texts could not be accurately identified, and therefore these programs could not identify all of the occurrences of some linguistic features. The programs have been designed to avoid skewing the frequency counts of features in one genre or another so that the relative frequencies were accurate. The main disadvantage of this earlier approach was that certain linguistic features could not be counted at all. For example, there was no way to compute a simple frequency count for the total nouns in a text, because nouns could not be identified. For these reasons, the second set of programs has been taking place.

The second stage of program development took place during the years (1985-1986). The approach used in this stage is different from that of the first stage. As a result, a general tagging program to identify the grammatical category of each word in a text was developed. The aim is to develop a program that was general enough to be used for tagging both written and spoken texts. For example, the program could not depend on upper case letters or sentence punctuation. This goal is achieved by using a large-scale dictionary together with a number of context-dependent disambiguating algorithms. The main problem that had to be solved is that many of the common words in English are ambiguous as to their grammatical category. Words like "absent" can be either adjectives or verbs; words like "acid" can be either nouns or adjectives. All past and present participial forms can function as noun (gerund), adjective, or verb. A simple word like that can function as a demonstrative, demonstrative pronoun, relative pronoun, complementizer, or adverbial subordinator.

[10] has developed algorithms to disambiguate occurrences of certain words, depending on their surrounding contexts. For example, a participial form preceded by an article, demonstrative, quantifier, numeral, adjective, or possessive pronoun is functioning as a noun or adjective. That is to say, it is not functioning as a verb in this context; given this preceding context, if the form is followed by a noun or adjective then it will be tagged as an adjective; if it is followed by a verb or preposition, then it will be tagged as a noun. Tagged texts enable automatic identification of a broad range of linguistic features that are major for differentiating between genres in English. The tagged texts are subsequently used as input to other programs that count the frequencies of certain tagged items (e.g. nouns, adjectives, adverbs) and compute the frequencies of particular syntactic constructions (e.g. relativization on subject versus non-subject position). This approach assures a higher degree of accuracy and it allows inclusion of some features that could not be accurately identified by the previous programs. The resulting analysis is thus more complete than earlier analyses. The researcher has consulted the IT team in *Stanford* who helped a lot in solving many problems that the researcher has encountered in this research. They also put Biber's algorithm into consideration, which is going to be an asset to their program.

2 METHODOLOGY

Biber is the main model upon which this study is based [10] . The initial step is to collect the English and Arabic texts that are used as the corpus of this study. The second step is to choose an English tagger to be implemented in the analyses of the data. A pilot study is conducted for choosing the best tagger to be used in the study which is Stanford tagger. Next, computational identification of the specified linguistic features in English and Arabic texts by the use of Stanford tagger is applied by the help of Stanford computer program which is computerized to do so. Furthermore, an annotation of the linguistic features specified by Biber in his model is manifested as the units of annotation and the choice of the annotation scheme are crucial for the equality of this research. Moreover, gathering the linguistic features into groups that occur with a high frequency is manifested. The following step is to group these features into factors and to apply a statistical analysis to interpret the features underlying each factor. Also, specifying the English and the Arabic features used in *Stanford* Corpus to be applied in both the English texts and the Arabic equivalent translated ones. Next, annotating the English / Arabic texts by using an English tagger, Stanford is applied. Text normalization is crucial for any comparison of frequency counts across texts, as text length can vary widely. A comparison of non-normalized counts is going to give an inaccurate assessment of the frequency distribution in texts. Finally, the actual

presence of the variables located in the texts and their parallel translated words are going to be checked in the SPSS program to calculate their actual number of presence.

The present study is significant for many reasons. First, the use of computer-based text corpora provides a standardized database. Second, the use of *Stanford computer program* to count the frequency of occurrence of 66 linguistic features in ten short stories and their translations and to offer a detailed analysis of the distribution of these features. It is the only computer program that can deal with the English and Arabic languages simultaneously. Next, the employment of multivariate statistical techniques, especially factor analysis, is applied to determine the co-occurrence relations among the linguistic features. Finally, the use of microscopic analysis is maintained to interpret the features underlying each factor.

While the purpose of this paper is academic, the need to accelerate the investigations in translation research is becoming a must, as translating from other languages in the modern era, with information flooding from every comer in the globe is increasingly in demand. Translation studies, have only evolved during the last decades [11]. Scientific research in this area is a very recent phenomenon, as stressed by [12]. The call for research in translation is overwhelming as "a whole range of issues seemed to be waiting for examination, and inquiry is overdue", [13]. Calls for conducting systematic comparative studies of translated and source texts [11] and those for research focusing more upon what [14] has termed "the acquisition of translation competence", have not been accomplished. In translating between English and Arabic, there is a shortage of research in translation problems that may be encountered by Arabic translators of English [15], [1]. T is presented ten short stories taken from Monro's collection of short stories *The Power of Love* and their translated equivalence which are translated by Mohamed Saad Tantawi and published by *Hindawi Foundation for Education and Culture* in 2015. The actual presence of the variables located in the texts and their parallel translated words are going to be checked in the SPSS program to calculate their actual number of occurrence. In order to normalize texts, in this study, the frequency counts of all linguistic features are normalized to a text length of 7,725 words so we have to delete some words to make them the same length.

3 FACTOR ANALYSIS: TECHNICAL DESCRIPTION

The first step in a factor analysis is to choose a method for extracting the factors. There are several options available to extract the factors, but the most widely used is known as ‘common factor analysis' or ‘principal factor analysis'. For instance, this procedure extracts the maximum amount of shared variance among the variables for each factor. Thus, the first factor extracts the maximum amount of shared variance. The second factor then extracts the maximum amount of shared variance from the tokens left over after the first factor has been extracted, and so on. In this way, each factor is extracted so that it is uncorrelated with the other factors.

Once a method of extraction has been chosen, the best number of factors in a solution must be determined, [9]. As noted above, the purpose of factor analysis is to reduce the number of observed variables to a relatively small number of underlying constructs. A factor analysis will continue extracting factors until all of the shared variances among the variables have been accounted for, but only the first few factors are likely to account for a nontrivial amount of shared variance and therefore be worth further consideration. There is no mathematically exact method for determining the number of factors to be extracted. A "scree plot", will normally show a characteristic break indicating the point at which additional factors contribute little to the overall analysis. The scree plot corresponding to eigenvalues is given in Figure (1), (2). The eigenvalues of the English and Arabic texts can be used to indicate the percentage of shared variance that is accounted for by each factor.

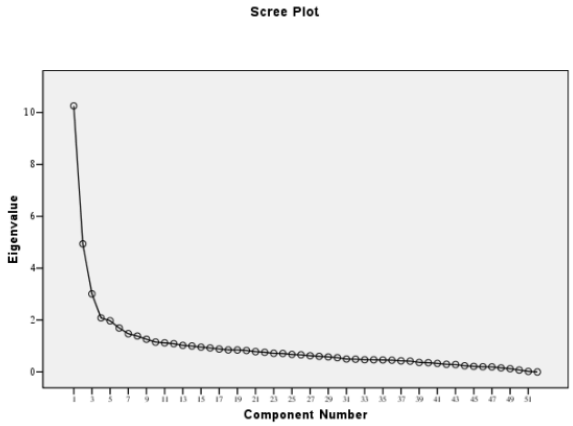

Figure 1: Scree plot of the English factors

The break in the English plot occurs between the first, second, third and fourth factors. When faced with a choice between a larger or smaller number of factors, the more conservative procedure is to extract the larger number and then discard any unnecessary factors [9]. Extracting too few factors will result in loss of information, because the constructs underlying the excluded factors will be overlooked; it might also distort the factorial structure of the remaining factors, because multiple constructs are collapsed into a single factor. The same procedure is applied to the Arabic data. The scree plot is applied to extract the number of factors needed and the features that constitute each factor.

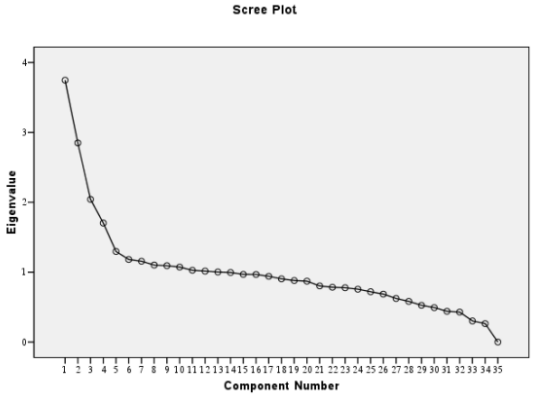

Figure (2) The Arabic scree plot.

The break in the Arabic plot occurs between the first, second, third, fourth and fifth factors. When faced with a choice between a larger or smaller number of factors, the more conservative procedure is to extract the larger number and then discard any unnecessary factors [9]. Extracting too few factors will result in loss of information, because the constructs underlying the excluded factors will be overlooked. It might also distort the factorial structure of the remaining factors, because multiple constructs are collapsed into a single factor. Factor loadings reflect the extent in which one can generalize from a given factor to an individual linguistic feature. Features with higher loadings on a factor are more representatives of the dimension underlying the factor, and when interpreting the nature of a factor, the features with large loadings are given priority. Multivariate statistical techniques such as factor analysis are not practical without the aid of computers. A factor analysis involves many computations using matrix algebra. The first point for a factor analysis is a simple correlation matrix of all variables. Factor analysis routines are usually included as part of the standard statistical packages (e.g. SPSS) available on computers at most academic institutions. SPSS computational tool makes a new range of linguistic research possible.

Factor analysis uses frequency counts of linguistic features to locate sets of features that co-occur in texts. The use of this technique to identify underlying textual dimensions is based on the assumption that frequently co-occurring linguistic features

have at least one shared function [10]. It is claimed here that there are relatively few primary linguistic functions in English and Arabic, and that the frequent co-occurrence of a group of linguistic features in texts is indicative of an underlying function shared by those features. Working from this assumption, it is possible to obtain a unified dimension underlying each set of co-occurring linguistic features. In this proposal, a collection of short stories written by Monro and their translation are used as the sample of this study. The next step is to identify the linguistic features with these texts before applying factor analysis. Inadequate preparation or skewing in these theoretical prerequisites can invalidate the results of a factor analysis [9]. That is, factor analysis provides the primary analytical tool, but it is dependent on the theoretical foundation provided by an adequate database of texts and inclusion of multiple linguistic features.

4 RESULTS

A. Interpretation of the English Results

The results of the present study reflect nine factors, four factors for the English language and five factors for the Arabic one. The nine factors identified here are general, underlying parameters of variations. Some of the features in the Arabic language cannot be identified in the five factors mentioned for the Arabic language. That is why the researcher has counted them manually by herself in a separate table. The factors do not represent all of the differences defined by the original 67 linguistic features that are identified by Biber in his model. The factors are abstractions, describing the underlying parameters of variations in relatively global terms.

Rotated component matrix in factor 1

	Factor 1
Positive	
Wh question	0.75
Type token ratio	0.72
Existential there	0.78
Amplifiers	0.47
Private verbs	0.64
Hedges	0.39
Contractions	0.72
Do as a pro verb	0..36
Word length	0.33
Past tense verbs	0.54
2nd persons pronouns	0.47
Pronoun it	0.45
Analytic negation	0.40
1st person pronoun	0.78
Wh relative clause subject position	0.45
Split infinitives	0.39
WH relative cl. in object position	0.50
Non phrasal coordination	0.40
Demonstrative	0.63
Emphatics	0.60
Negative	
Present tense	-0.07
Attributive adjective	-0.08

t, p: t and p values for Student t-test

*: Statistically significant at $p \leq 0.05$

Rotated component matrix in factor 2

	Factor 2
Past participial clause	0.52
Agentless passive	0.66
By passive	0.78
Nouns	0.83
Present tense verbs	0.73
3rd person pronoun	0.51
Perfect aspects	0.62
Public verbs	0.77
Synthetic negation	0.59
Present participial clause	0.77
Attributive adjectives	0.89
Present participial WHIZ deletion	0.42
Prepositional phrase	0.69
That deletion	0.50
Conjuncts	0.64
Adverbs	0.67
Negative	
Word length	*-0.58*
Prepositions	*-0.54*

t, p: t and p values for **Student t-test**

Rotated component matrix in factor 3

	Factor 3
Predicative adjectives	0.43
Other adverbial subordinators	0.75
Gerunds	0.81
Time adverbials	0.40
Place adverbial	0.50
Adverbs	0.67
Be as a main verb	0.82
Pied piping construction	0.63
Prediction modals	0.50
Conditional subordination	0.80
Discourse particle	0.50
Possibility modals	0.63
Necessity modals	0.73
Suasive verbs	0.48
Consessive subordination	0.67

Rotated component matrix in factor 4

	Factor 4
Positive	
That clause as a verb compliment	0.44
Past participial WHIZ deletion	0.58
That clause as adjective compliment	0.39
That clause on subject position	0.54
Sentence relatives	0.71
Demonstrative pronouns	0.63
Indefinite pronoun	0.53
Seem / appear	0.61
Down toners	0.38
That clause on object position	0.67
Nominalization	0.83
Causative subordination	0.48
Split auxiliaries	0.25
Infinitives	0.60
Phrasal coordination	0.67
Demonstrative pronoun	0.63
Negative	
Time adverbial	-0.50
Place adverbial	-0.49

t, p: t and p values for Student t-test

*: Statistically significant at $p \leq 0.05$

Overall, these results indicate that the tagging program is quite accurate as some of the tagged items were counted manually to reassure the accuracy of the program. First, there are very few mis tags; the majority of 'errors' are untagged items, which do not introduce misleading analyses, and even untagged items are relatively uncommon. Secondly, there is no serious skewing of mis tags so that the results are accurate in relative terms; that is, the results enable accurate comparisons across texts because the same word types are left untagged in all texts. Last but not least, the few mistags and untagged items that do exist are of a very specialized or idiosyncratic in nature, and often these items have no bearing on the linguistic features counted for the analysis of textual dimensions. The tagged texts produced by this program thus provide a good basis for the automatic identification of the linguistic features, only the potentially important linguistic features are actually counted.

The tagging of some lexical items was so problematic that they were systematically excluded. In addition, the researcher has carried out some hand editing of the tagged texts to correct certain inaccuracies. For example, past and present participial forms were checked by hand. Although the tagging program includes elaborate algorithms to distinguish among gerunds, participial adjectives, WHIZ deletions, participial clauses, passives and perfects (in the case of past participles), and main active verbs (present or past), a high percentage of these forms was incorrectly tagged.

To a computer program without access to semantic information, however, there is no difference between these constructions, and thus at least one of the two cases will be tagged incorrectly. Similar problems were found in attempting to disambiguate the other functions of present and past participial forms. As a result, all participial forms were checked by hand. The factors reflect the fact that depictive details are important in narrative discourse. Discourse particles are generalized markers of informational relations in a text. They help to maintain textual coherence when a text is fragmented and would otherwise be relatively in coherent. Also, subordination features occur with a variety of involved and generalized content features, and in a complementary pattern to highly informational features. Furthermore, sentence relatives are present to express attitudinal comments. Wh – clauses provide a way to "talk about" questions. Time and place adverbials depend on referential inferences by the addressee.

Persuasion is one of the main techniques used in informative texts; it is a marking of the author's own point of view or an assessment of the advisability of an event presented to persuade the reader. Narrative genre is marked by considerable reference to past time, third person animate referents, reported speech and depictive details. It has a high lexical variety. It is a discourse that reports events in the past or deals with more immediate matters but does not mix both. In conclusion, the four factors have strong factorial structures and the features grouped in each factor are functionally coherent and can be easily interpreted.

B. Interpretation of the Arabic Factors

Rotated component matrix in factor 1

	Factor1
Amplifiers	0.51
Analytic negation	0.78
Conditional subordination	0.66
Discourse particles	0.57
Emphatics	0.74
Pied piping	0.30
Prepositional phrase	0.61
Private verbs	0.66
Seem and appear	0.73
Wh clause	0.67
Type/ token ratio	0.71
Sentence relatives	0.64
Split auxiliary	0.54
Infinitives	0.50
Causative subordination	0.53

Rotated Component Matrix in factor 2

	Factor 2
Place adverbial	0.53
Present participial WHIZ deletion	0.39
Present participial clause	0.87
Public verbs	0.76
Adverbial past participle clause	0.37
Demonstrative pronouns	0.37
Indefinite pronoun	0.44
Past tense	0.87
Perfect aspect verbs	0.73
Synthetic negation	0.42
That clause on object position	0.70
Negative	
Past participle WHIZ deletion	-0.35

Rotated component matrix in factor 3

	Factor 3
Past part. Clause	0.43
Present tense	0.80
3rd person pronoun	0.71
Adverbs	0.40
Gerunds	0.40
That clause as adjective compliment	0.46
That clause an subject position	0.54
Other adverbial subordinators	0.49
Time adverbial	0.29
Attributive adjectives	0.70
Negative	
Past tense	-0.45

t, p: t and p values for **Student t-test**

*: Statistically significant at $p \leq 0.05$

Rotated component matrix in factor 4

	Fctor4
Phrasal coordination	0.62
Nominalization	0.81
Conjuncts	0.51
Existential there	0.48
Hedges	0.62
Wh relative clause on object position	0.47
Wh relative clause on subject position	0.45
That clause as a verb compliment	0.61
Negative	
Synthetic negation	-0.38

t, p: t and p values for **Student t-test**

*: Statistically significant at $p \leq 0.05$

Rotated component matrix in factor 5

	Factor 5
Suasive verbs	0.50
Word length	0.36
1st person	0.29
Demonstratives	0.87
Non phrasal coordination	0.47
2nd person	0.34
Nouns	0.45
Past participle WHIZ deletion	0.50
Concessive subordination	0.47
Negative	
Nominalization	-0.36
Time adverbials	-0.31
3rd person pronoun	-0.39

B.1 Features not in the Arabic Results

Features not in the Arabic Results	No. English	No. Arabic (manual)
Pronoun it	3705	3705
Be as a main verb	14840	5687
Do as a pro verb	19400	-
Subordination that deletion	57113	-
Split infinitives	5567	1000
Possibility modals	20300	20300
Necessity modals	5950	5950
Prediction modals	20300	20300
Contractions	15100	-
By passive	20276	706
Agentless passive	8167	1700
Predicative adjective	36033	1800

The coming section deals with the features that *Stanford* program could not identify in the Arabic language .All the features encountered in this table are counted manually by the researcher herself; only 12 features are counted manually and this is done because the computer program, Stanford, could not identify these features due to the complex nature of the Arabic language.

The first feature is the modal verbs. Necessity and possibility modals are used either as an explicit marking of the writer's own point of view or as an argumentative discourse designed to persuade the addressee. In addition, the necessity modals are pronouncements concerning the necessity of certain events and the possibility modals are pronouncements concerning the possibility of certain events occurring. Suasive verbs and conditional subordination act as an alternative for the prediction modals

in Arabic. They imply intensions to bring about certain events in the future while conditional subordination specifies the conditions that are required for certain events to occur.

For example:

1-But knew I shouldn't waste the milk.

كنت اعلم انني لا يجب ان اهدر اللبن

2-He should have not stayed.

كان لا ينبغي عليه البقاء

The above examples illustrate the variety of positions in which the English negation can occur. In all cases, the Arabic counterpart immediately precedes the verb. English usage will seem extremely random and complex to the Arabic-speaking student. Modals present a variety of problems to the Arabic translators of English since modals as grammatical classes do not exist in Arabic. Their meanings are conveyed by particles, prepositional phrases, and unmodified verbs.

For example:

3-She can use the telephone.

كان في استطاعتها ان تستخدم التليفون

Moreover, "can" can be rendered in Arabic as a prepositional phrase. In most cases, such a verb or prepositional phrase precedes a nominalized /?an/ clause.

Arabic translators are not familiar with vowel reduction as it occurs in English, and are likely to use the full form in all cases as the Arabic language does not allow the contraction technique in its characteristics except in some forms as:

4 - صلى الله عليه وسلم (ص)

As for the passives and other past participial clauses, they are used to emphasize abstract conceptual information over more concrete or active content. Usage of the passive form with (was/ were/ is /are) in English is totally neglected in the Arabic translation. The passive form does not exist in the Arabic language. It is not used as we drop it in the translation of the Arabic language.

For example:

5- She was cut down and taken away.

تم انزالها وحملت للداخل

Verb to be in "taken away" has been omitted by the translator and has been used only the past participle of the verb as it has been used in the first verb (was cut).

6- People are dead now.

انهما ميتان الان

The translator avoided the passive structure to change the sentence into a nominal sentence and he omitted verb to be from the sentence. He also translated /people/ as /انهما/

7- I was struck by that.

ادهشني ذلك

The translator changed the passive sentence into an active one so it can be more vivid to the reader.

8- My brothers were not bothered by any of this.

لم يكن أخواي ينزعجان

The translator changed the passive sentence into an active one. He used /lam/ as a particle to change the verb in the present to give the meaning of the past. Predicative adjectives are proceeded by a linking verb to be. Some verbs have a verb (1) and verb (2) forms. Verb (1) being an intransitive linking verb meaning seemed or appeared followed by a predicative adjective.

For example

9- Father was polite.

كان والدي غايه في التهذيب

10- That kind of life is dreary.

هذا نوع كئيب من الحياه

The verb to "be" in the present tense is neither used in the Arabic language nor verb to do. Verb to do expresses about the past simple tense /did/ or it expresses about the absence of the third person /does/

For example

11- He does not care.

لا يهتم

In this sentence "does" is used because of the word "he" and to remove the "s" of the verb to express the continuous tense with the pronoun.

12- I figured out that he didn't mind people doing new sorts.

ادركت انه كان لا يكترث بقيام الناس بأشياء جديدة

Here "did" is used to express that the verb is in the past.

That deletion, while that can be dropped in English, /?anna/ should be always retained in Arabic.

For example:

13- Just what is it you are famous for?

كنت اود ان اسألك: بم انت مشهوره؟

"That" is dropped in the sentence but /?anna/ is not. Arabic uses relative nouns that need to agree with the head noun in case, gender, and number. The verb to "be" in the present tense is neither used in the Arabic language nor verb to do. Verb to do expresses about the past simple tense /did/ or it expresses about the absence of the third person /does/

[16] has stated that there is no neuter pronoun in Arabic, that is to say, pronoun "it". It uses only feminine and masculine gender. The English impersonal it has no counterpart in Arabic. These independent pronouns as claimed by [17] can function as a subject of a verb, a subject or a predicate of a verb less sentence and as a copula. The English pronoun it has no counterpart in Arabic. For example:

14- They washed and combed it beautifully.

وكن يغسلن و يمشطن شعره علي نحو رائع

15- She could pay it back.

حتي يستطيع دفع ما يقابلها له

16- It was raining outside.

كانت تمطر بالخارج

As shown in these examples pronoun “it” is translated like the 3rd person pronoun because of the nature of the language itself. That is why the 3rd person pronoun is significantly larger in weight in Arabic than English. Concerning “be” as a main verb, it is typically used to modify a noun with a predicative expression instead of integrating the information into the noun phrase itself. Be as a main verb is omitted in the translation thus changing the English verbal sentence into Arabic nominal ones. That is to say into a topic and a comment. When (am, is, are) are used as main verbs, their sentences are nominal in Arabic. Therefore, they are deleted completely in Arabic. The past tense of (be, have) are translated into verbal sentence in Arabic and this is more effective in delivering the message. For example:

17- My father was not religious.

لم يكن ابي متدينا

In other cases the translator needs to change the phrase to verbal sentences and to remove verb to be.

18- At the end of the yard is a small barn.

يوجد في نهاية الفناء مخزن صغير

There is another difference in translating the following sentence:

19-They are dead now

انهما ميتان الآن

It is a passive sentence and that is why here in translating passive sentences we can use nominal sentences affirmed with "ان" . In a sentence like:

20- Both of my boys were in school.

ولداي يذهبان إلى المدرسة

In this example verb to be is omitted and the word/ both/ is deleted and the duality is shown in the word "كلتا". We can conclude here that it shows that the plural in Arabic is changed into dual.

5 CONCLUSION

In view of the paper presented here, linguistic variations are considered as a field of study that requires further analyses based on the use of corpora and the refinement of parameters in register description. Obviously, register variation research has immediate applications to foreign language teaching and intercultural communication, and this type of perspective that the field offers should attract scholars and communication practitioners. Biber’s model is only dealing with the morphology and the syntax of the language. More models are needed to combine the structure and the ideology together. This sort of descriptive study is greatly facilitated by the availability of tools of corpus linguistics. The *Stanford* program used in this investigation is user-friendly and has proved very practical as an aid to human analysis of a whole text. The tagging could be grammatical (to look more closely at clause beginnings or shifts from noun to verb), functional (such as analysis of Transitivity patterns) or stylistic (the highlighting of the occurrence of particular lexical fields, an author's favorite constructions, words with positive and negative connotations.

REFERENCES

[1] Khafaji, R. (1996). Arabic translation alternatives for the passive in English, Papers and Studies in Contrastive Linguistics, 31,19-37.

[2] Buránová, E., Cová, E. & Sgall, P. (2000), “Tagging of very large corpora: Topic-focus articulation”. In Proceedings of the 18th conference on Computational Linguistics (Coling), 1, 139–144, Saarbrücken: Germany.

[3] Olohan, M. (2004). Introducing corpora in translation studies. Taylor and Fracnis: London.

[4] Teubert, W. (2002), "The Role of Parallel Corpora in Translation and Multilinguial Lexicography". In Lexis in Contrast, B. Altenberg and S. Granger (eds.), 189 – 214. Amsterdam: Benjamins.

[5] Mitkov, R (ed.). (2015). The Oxford hand book of computational linguistics. 1st ed. Oxford University Press: Oxford

[6] Fisiak, J. (1981). Some introductory notes concerning contrastive linguistics. In J. Fisiak (Ed.), Contrastive linguistics and the Language Teacher (pp. 1-11). Oxford: Pergamon.

[7] Towell, R. & Hawkins, R. (1994). Approaches to second language acquisition. Clevedon: Multilingual Matters.

[8] Crystal, D. (1992). Introduction to basic linguistics. England: Penguin Books Ltd

[9] Gorsuch, R. (1983). Factor Analysis (2nd Ed.). Hillsdale, NJ: Erlbaum.

[10] Biber, D. (1988). Variation across speech and writing, Cambridge: Cambridge University Press.

[11] Broeck, R. (1986). Contrastive discourse analysis as a tool for the interpretation of shifts in translated texts. In House, J & Blum Kualka, S (Eds.) Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies, Tubingen: Gunter Narr PP. 37-47.

[12] Gile, D. (1994) Beyond testing towards a theory of educational assessment, London: Falmer Press.

[13] Simon, S. (1996). Gender in Translation: Cultural Identity and politics of Translation. London and New York: Routledge.

[14] Krings, H. (1986). Translation problems and translation strategies of advanced German learners of French (L2). In House, J. & Blum-Kulka, S. Interlingual and Intercultural Communication: Discourse and Cognition in Transition and Second Language Acquisition Studies. Tubingen: Gunter Narr Verlag, 263-276.

[15] Khalil, A. (1993). Arabic translations of English passive sentences: problems and acceptability Judgements, Papers and Studies in Contrastive Linguistics, 27, 169-81.

[16] Affendi, A. (2011). A contrastive analysis between Arabic and English relative pronouns. Faculty of Education: Salatiga. Retrieved from https://docplayer.net/39367735.

[17] Ryding, K. (2005). A reference grammar of modern standard Arabic. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511486975.

BIOGRAPHY

Maram Elsaadany

She has attained her PH-D degree in Translation studies from the Institute of Applied Linguistics and Translation (2018), Faculty of Arts, Alexandria University. She has also attained her Masters degree in Applied Linguistics in (2014). Her main areas of interest are applied linguistics, computational linguistics and computational studies in the field of translation. She is also the head of an educational center for teaching the CIPP program.

Elsaadany is a former English instructor in the Arab Academy for Science and Technology. Furthermore, she is a speaking examiner for IELTS and OET exams**.**

Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Translated Abstract

مقترح أداه لقياس التغيرات اللغوية: دراسه مؤسسه علي مدونه لغويه

مرام السعدني

معهد الدراسات اللغوية و الترجمة – كليه الاداب- جامعه الاسكندريه – مصر

Maram.elsaadany@gmail.com

سامح الانصاري

قسم الصوتيات –كليه الاداب – جامعه الاسكندريه

sameh.alansary@bibalex.org

ملخص :

هذا البحث هوتحليل لترجمة للهياكل المورفولوجية العربية والإنجليزية باستخدام نموذج بيير (1988) وبرنامج ستانفورد وهي عبارة عن دراسة كمية تستند إلى مجموعة من الخصائص اللغويه و تستخدم 66 ميزة من بين 67 سمة تم تحديدها من قِبل بيير ، ويتماشى البحث مع نموذج بيير الإجراء و الإحصائي. إن المجموعة المختارة لهذه الرسالة هي مجموعة قصص أليس مونرو القصيرة The Power of Love (1985) وترجمتها إلى العربية مسيره الحب (2015) لمحمد طنطاوي. يتم حساب جميع االتغيرات اللغويه في اللغه الإنجليزية والعربية بمساعدة برنامج كمبيوتر ، ستانفورد (2015). ستانفورد هو برنامج يستخدم لتلخيص مجموعة النصوص المختارة ، وهو يعمل على مستويات اللغة الإنجليزية والعربية .morpho-syntactic يتم تصنيف 66 وظيفة إلى أربعة عوامل للغة الإنجليزية وخمسة عوامل للغه العربية. يتم حساب اثني عشر وظيفة يدوياً في التحليل العربي بواسطة الباحثه بنفسها. تعكس النتائج اختلافات كبيرة بين اللغتين حيث لا يمكن تحديد بعض الميزات بواسطة برنامج الكمبيوتر.

الكلمات المفتاحية: الترجمه / حسابي / اختلافات لغويه