مجلة جامعة بابل للعلوم الإنسانية، المجلد ٢٦، العدد ٨: ٢٠١٨.
Journal of University of Babylon for Humanities, Vol.(26), No.(8): 2018.

Corpus Linguistics and Translation

Riyadh Khalil Ibrahim
Al-Esraa University College
drtememe@ gmail.com

Submission date:25 /4/2018 **Acceptance date**: 9/9/2018 **Publication date**: 19/11 /2018

Abstract

The 21st century witnesses tremendous technological and organizational advances in world's economy and societies. This has left a great impact on translation and translation studies. Corpus is a machine-readable representative collection of naturally occurring language assembled for linguistic analysis accessible with software such as, concordances that can find list and source of linguistic patterns. It lays the foundation of Corpus linguistics which makes it possible for translators and translation studies to make use of large quantity of stored data on computers for examining target language translations. Computer corpora includes spoken/written, casual/formal, fiction/non-fiction texts representing various demographic areas. The study aims at familiarizing student of translation and translators with the methods and practical applications of computer corpora in various fields of language use. The study reveals that Corpus data are essential for accurately describing various samples of language by showing how lexis, grammar and semantics interact to serve appropriate translation output.

Key words : corpus linguistics, computer corpora, corpus- based translation.

علم لغة المدونات والترجمة

رياض خليل ابراهيم
كلية الاسراء الجامعة

الخلاصة

شهد القرن الواحد والعشرين تطورا تكنولوجيا وتنظيميا في الاقتصاد والمجتمع الدولي مما عكس أثرا كبيرا على الترجمة والدراسات الترجمة. تعد المدونة مجموعة لغوية طبيعية تعرض للقراءة على الحاسوب لاغراض التحليل اللغوي لنماذج لغوية متنوعة وهي الأساس لظهور علم لغة المدونات. لقد عمل علم لغة المدونات على تسهيل مهمة المترجمين في الاستفادة من كميات هائلة من المعلومات المخزونة في الحاسوب لغرض دراسة وترجمة اللغة الهدف . تهدف الدراسة الحالية الى تعريف طلبة الترجمة والمترجمين على طرق التطبيق العملي لمدونات الحاسوب لاستخدامها في مختلف حقول المعرفة. لقد أظهرت الدراسة أن علم لغة المدونات ذا أهمية بالغة في التوصيف الدقيق لنماذج مختلفة من اللغة حيث يوضح كيفية التفاعل بين المفردات والنحو وعلم لغة الدلالة لإنتاج الترجمة الملائمة. تضم مدونات الحاسوب نصوصا مكتوبة، شفوية، عامية، رسمية، روائية وغير روائية لمختلف المناطق السكانية وعلم جغرافية اللغة ومن هنا تبرز الحاجة لتكوين مدونات عربية وطنية للمؤلفين العرب لغرض المقارنة مع المدونات الاجنبية.

الكلمات الدالة: علم لغة المدونات، مدونات الحاسوب، الترجمه القائمة على المدونات.

1. Introduction

Corpus linguistics has become a major paradigm and research methodology in translation theory and practice, with practical applications ranging from professional human translation to machine (assisted) translation and terminology. Corpus-based theoretical and descriptive research has investigated written and interpreted language and topics such as translation universals and norms, ideology and individual translator style. In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of text usually electronically stored and processed. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within specific language territory. Corpus linguists have been logging occurrence and distributions of language forms in search of usage patterns that identify different registers and help understand the contextual factors which influence its variability. The flowering of corpus linguistics inspired a similar empirical turn in translation studies. In the first two computer-aided decades, computer-based translation studies has initially focused on product rather than process, exploring the characteristics of translations as texts, or as pattering specific to translational language. When a new technologies are first introduced into a profession, it is not possible to know immediately what the long- term effect will be. It is only with time that the impact of applying a given technology can be observed. Electronic corpora and associated tools for corpus processing have now been in widespread use in the translation profession for approximately fifteen years [1:7].

Corpus is often defined as a machine-readable representative collection of naturally occurring language assembled for the purpose of linguistic analysis. Its analysis lays the foundations for corpus linguistics. Corpus linguistics is not a homogeneous methodology: it is used with a varying level of granularity and varying reliance on quantitative and qualitative methods, with its shared features being as follows: machine-readable naturally-occurring language, balanced and representative corpus design, systematic and exhaustive analysis.

- the analysis is based on a corpus or corpora of naturally-occurring languages which are machine-readable so that the retrieval of the search patterns is computerized.
- the corpus is intended or taken to be balanced and/or representative of the modality/register/variety the study is aimed at.
- the analysis is, or at least attempts to be, systematic and exhaustive, meaning that the corpus does not simply serve as a database of examples from which some can be chosen *ad libitum* and others neglected but that the whole (sample of the) corpus is taken into consideration[2:4].

2.Language Corpora.

According to[3 :110], a 'language corpus' usually means a text collection which is of different types as follows:

1. large: millions, or even hundreds of millions, of running words, usually sampled from hundreds or thousands of individual texts.
2. computer-readable: accessible with software such as concordances, which can find, list and sort linguistic patterns.
3. designed for linguistic analysis: selected according to a sociolinguistic theory of language variation, to provide a sample of specific text-types or a broad and balanced sample of a language.

Much 'corpus linguistics' is driven purely by curiosity. It aims to improve language description and theory, and the task for applied linguistics is to assess the relevance of this work to practical applications. Corpus data are essential for

accurately describing language use, and have shown how lexis, grammar and semantics interact. This in turn has applications in language teaching, translation, forensic linguistics, and broader cultural analysis. In limited cases, applications can be direct, for example, if advanced language learners have access to a corpus, they can study for themselves how a word or grammatical construction is typically used in authentic data. However, applications are usually indirect. Corpora provide observable evidence about language use, which leads to new descriptions, which in turn are embodied in dictionaries, grammars and teaching materials. [3:112]

Modern computer-assisted corpus study is based on two principles:

1- The observer must not influence what is observed.
   What is selected for observation depends on convenience, interests and hypotheses, but corpus data are part of natural language use, and not produced for purposes of linguistic analysis.
2- Repeated events are significant.
   Quantitative work with large corpora reveals what is central and typical, normal and expected. Corpus study is inherently sociolinguistic, since the data are authentic acts of communication; inherently diachronic, since the data are what has frequently occurred in the past; and inherently quantitative. This disposes of the frequent confusion that corpus study is concerned with simple performance pejorative sense of being characterized by memory limitations, distractions, shifts of attention and interest, and errors. The aim is not to study idiosyncratic details of performance which are, by chance, recorded in a corpus. On the contrary, a corpus reveals what frequently recurs, sometimes hundreds or thousands of times, and cannot possibly be due to chance.

3.Available Corpora

According to [4:41] contend that any list of extant corpora would be quickly out of date, but there are two sets of important distinctions between:

1. small first generation corpora from the 1960s onward and much larger corpora from the 1990s, and
2. carefully designed reference corpora, small and large, and other specialized corpora, opportunistic text collections, archives and the like.

The first computer-readable corpora, compiled in the 1960s, are very small by contemporary standards, but still useful because of their careful design. The Brown corpus (from Brown University in the USA) is one million words of written American English, sampled from texts published in 1961: both informative prose, from different text-types (e.g., press and academic writing), and different topics (e.g., religion and hobbies); and imaginative prose (e.g., detective fiction and romance). Parallel corpora were designed to enable comparative research: the LOB corpus (from the universities of Lancaster, Oslo,& Bergen) contains British data from 1961; Frown and FLOB (from Freiburg University, Germany) contain American and British data from.

International Corpora of English(ICE) contains regional varieties of English, such as Indian and Australian. Similar design principles underlie the Lund corpus of spoken British English (from University College London and Lund University), which contains around half a million words, divided into samples of the usage of adult, educated, professional people, including face-to-face and telephone conversations, lectures and discussions. By the late 1990s, some corpora consisted of hundreds of millions of words. The Bank of English (at COBUILD in Birmingham, UK) and the British National Corpus (BNC) had commercial backing from publishers, who have used the corpora to produce dictionaries and grammars. The 100-million-word BNC is also carefully designed to include demographically and stylistically

defined samples of written and spoken language. The Bank of English arguably over-emphasizes mass media texts, but these are very influential, and it still has a range of text-types and advantages of size: over 400 million words by 2001. Because constructing large reference corpora is so expensive, it may be that huge new corpora cannot again be created in the near future. These corpora will remain standard reference points, which can be supplemented by small specialized corpora, designed by individual researchers, and by large opportunistic collections of many other corpora for English, and increasingly for other languages [5:176].

4. Corpus Design and Methodology

Some basic principles of corpus design are simple enough. A corpus which claims to be a balanced sample of language use must represent variables of demography, style and topic, and must include texts which are spoken and written, casual and formal, fiction and nonfiction, which vary in level (e.g. popular and technical), age of audience (e.g. children or adults), and sex and geographical origin of author, and which illustrate a wide range of subject fields (e.g. natural and social sciences, commerce, and leisure). However, no corpus can truly represent a whole language, since no-one quite knows what should be represented. It is not even obvious what are appropriate proportions of mainstream text-types such as quality newspapers, literary classics and everyday conversation, much less text-types such as newspaper , business correspondence and church sermons.[6:93] .

Corpus design and methodology dates back to the 1980s. Yet this methodology was known much earlier in a paper form. Some linguists describe its 'Stone Age' and list a few examples of 'language corpora BC' (before computers), (6&3). The most notable one is a corpus of 5 million citation slips compiled by volunteers in the second half of the 19th c. and at the beginning of the 20th c. for the Oxford English Dictionary published in 1928 (see 3 : 110). Other famous linguists who used 'shoebox corpora' included: Jespersen, Boas, Sapir, Fries, Bloomfield, and Pike [4:66]. The fifties brought about a move away from empirical methods in favor of the rationalism paradigm, following Chomsky's interest in linguistic competence and related skepticism of corpora due to their 'skewedness'. Despite the unsupportive attitude of linguistic circles it was in the 1960s when the first modern machine-readable corpus, the Brown University Standard Corpus of Present-Day American English, was built by Henry Kučera and W. Nelson Francis. It had only a million words and at that time its processing required the application of all available computer resources [7:20]. Following developments in computer science, corpus-based studies of language re-emerged on a greater scale in the late 1980s, spreading into all possible areas and branches of linguistics and related disciplines. Its popularity may be confirmed by the sheer number of books and articles published on the topic. With time corpora have markedly increased in size: the Oxford English Corpus has 2 billion words, the Corpus of Contemporary American English has 400+ million words and the British National Corpus has 100+ million words.[1:22].

5.Features of Translation Studies

Communication between different individuals and nations is not always easy, especially when more than one language involved. Translation is undoubtedly a communicative device, which is described ," The fact is, translation is a necessity on economic and on general human grounds"[8:135]. Translation studies, as postulated by some researchers, brings together work in a wide variety of fields, including literary study, anthropology , psychology, and linguistics . Others claim that the domain translation studies is an important sub-branch of applied linguistics. Proponents of both opinions would have to admit that the field of translation studies

has multidisciplinary dimensions and aspects. The term" translation" refers to a written materials but is also used for all tasks where elements of a text of one language (the source language SL) are modeled into a text of another language (the target language TL), whether the medium is written, spoken, or signed. The job of the translator and/or interpreter is to try to bridge the gap between two foreign languages. Machine translation is a widespread and new way to have a complete and ready-made comprehensive translated texts[9:692].

Features of translation could be described into three classes according to their correlations, especially in their operationalization, as follows:

1- **Simplification** can be analyzed on different levels, e.g. lexical, syntactic or semantic. If core patterns of lexical use are observed, we can identify simplification comparing the proportion of content vs. grammatical words. Translated texts have a relatively low percentage of content words, and the most frequent words are repeated more often. This means, that both lexical density and type–token–ratio of translations are lower than those of their source texts and the comparable texts in the target language. Besides, more general terms are expected to be used in translations. On the level of syntax, one can observe short sentences which replace long ones and a lower average sentence length in general. ***Explication*** involves the addition and specification of lexical and grammatical items, with the help of which implicit information in the source text is "spelled out" in its translation. The indicators of this feature include a higher ratio of function words which make grammatical relations explicit, specific terms replacing more general terms (the opposite of simplification), disambiguation of pronouns, increased use of cohesive devices, e.g. conjunctions, and others. In terms of cohesion, one would also expect more nominal (expressed with nominal phrases) than pronominal reference (expressed with personal pronouns) in translations. Simplification and explication features correlate and may be just the opposite of each other. For example, if we observe more specific terms replacing general terms in translation, we face the feature of explication, and not simplification.

2- **Normalization** and "**shining through**" can also be measured on different levels, depending on the languages involved. Both features depend on the contrasts between these languages: ***normalization*** implies the exaggerated use of the patterns typical for the target languages, whereas "***shining through***" involves the variation in translation patterns typical for the source language (but not specific for the target language) that can be observed in translations. For instance, *normalization* can be verified by a great number of typical collocations and neutralized metaphoric expressions. The influence of *normalization* depends on the status of the source language: "the higher the status of the source text and language, the less the tendency to normalize". We assume that the languages with a higher status also tend to "*shine through*" more often. For example, if we analyze translations from English, we would probably observe more "shining through" than normalization, as English has the highest world language status[5:180[.

3- **Convergence** is a homogeneity feature of translations: they reveal less variation if we compare them to original texts. Convergence can also be observed on all levels of a language system. In accordance with the convergence phenomenon, one would expect that the lexical, grammatical and syntactic features under analysis will reveal smaller differences in translations than in originals.[6:96-97].

Much valuable work has been carried out by "translatologists" on the methodical scrutiny of translation, establishing interesting but often contradictory translation principles. Some principles run as follows:

1.a translation must give the words of original.
2.a translation must give the idea of original.
3.a translation should read like an original work.
4.a translation should read like translation.
5.a translation may add or omit from the original.
6.a translation may never add or omit from the original [10:54].

6.Machine Translation and Computer-assisted Translation

Machine-aided human translation (MAHT) ,also known as computer-assisted translation (CAT), involves some interaction between the translator and the machine must be distinguished from fully automatic machine translation(FAMT), better known as machine translation (MT), is characterized by the absence of any human intervention during the translation process. The purpose of a machine translation system is the same as that of any translation system: taking text written or spoken in one language and writing or speaking it in another one. Translation poses challenging problems both for the human translator and for the machine attempting to do what the human does. The machine- aided human translation approach seems to be more suited to the needs of many organizations which have to handle the translation of documents. Computer-assisted translation systems are based on translation memory. With such systems that sometimes combined with terminology database, translators have immediate access to previous translation of portions of the text, which they can then accept, reject, or modify. By constantly archiving their final choice, translators will soon have access to an enriched memory of ready-made solution for a wealth of translation problems. Other recent developments in computer technology also help the translators to perform their job. There is, for instance, a new and very affective productivity tool available for PC-based translators : automatic dictation software. At the present state of speech-recognition technology, however, to use dictation effectively the translator must master a new foreign language "Paused" speech that the computer can understand [9:702].

7. The Establishment of Corpora in Applied Translation Studies

Applied computer translation studies (CTS) took off slightly later compared with descriptive studies and then grew fairly rapidly. Its beginning can be traced back to [11:88] and [12:55],but it is really from the end of the 1990s onward that we can really talk of growing body of research in this area. Applied corpus studies of translation forged strong links with contrastive analysis, language for specific purposes (LSP), foreign language teaching terminology, lexicography and computational linguistics. At the core of corpus-based pedagogy were the design and navigation of corpora created not only as sources for the retrieval of translation equivalents or as aids for improving the quality and efficiency of the final translation product, but also as repositories of data used to better understand translation processes and language . So at the end of the 1990s the overall picture looks like this: within the empirical paradigm whose development in the early 1990s can be regarded as the most important trend that characterizes translation studies, a number of novel syntheses in the pure and applied branches of the discipline were proposed and realized with corpus linguistics methods.

According to[13:657], at the end of this initial period of intense scholarly work, three main areas of development for the new millennium can be identified :*First*, the search for common ground between linguistics and the rapidly developing

interdisciplinary field of cultural studies. *Second*, an awareness of ideology as a factor indissolubly intertwined with text, context, translation as well as the theory, practice and pedagogy of translation. *Third,* keeping pace with the development of modern technologies in order to continually update and refine the diversity of the methodologies in descriptive and applied studies[14:20].

The term corpus when used in the context of computer-based translation studies has more specific connotations than traditional definitions such as the one provided, for instance, by the Oxford Concise English Dictionary (i.e. "a large collection of written or spoken texts"), which does not carry such connotations. These connotations can be associated with at least four main attributes as follows:

1- *Electronic form* – for many years the word 'corpus' was only associated with hard-copy texts, but after the advent of the computer, it nearly always implies a collection of texts held in electronic form which can be read and analyzed automatically or semi-automatically rather than manually [15: 226].
2- *Size from*– a historical perspective, corpus-based studies have often relied on huge amounts of data in order to increase empirical evidence and knowledge about the world of our experience. As a consequence of this fact, the term "corpus" has traditionally been associated with vast quantities of data extracted from large collections of text; nevertheless, in the context of CTS the term has also been used to describe what came to be known as "small-scale corpora" in translator education. Therefore the issue of corpus size in CTS becomes a relative one in the sense that qualitative aspects sometimes may be more relevant than quantitative ones.

Another important aspect related to size has to do with the use of full texts instead of text fragments. Corpora which consist of full texts are by and large far more useful than those which consist of text fragments. This is so because full texts allow for the examination of not only micro level units such as words, phrases and sentences, but also the way texts are structured in their entirety, that is to say, how texts are formed by chapters, sections, paragraphs and so on [15: 225].

3- *Representativeness* – in building a corpus covering an area of interest, researchers must know to what extent and in what respects their corpus is representative enough to serve its purpose. Thus, the selection of texts in a representative corpus is not only related to size, but also to a careful description of what the corpus is intended to represent. Moreover, in the case of parallel corpora one has to establish unequivocally the source texts of the translations as there are times when "a multitude of candidates for a source text may exist". By so doing, researchers would not be faced with an injudicious choice of source text, which could certainly lead them astray and consequently produce rather unfortunate results.
4- *Open-endedness* – this refers to the flexibility that a corpus in translation studies should have to enable researchers to answer specific research questions. In other words, by means of an opened-ended corpus researchers can select and use the texts of this corpus for different types of comparisons and studies. Therefore, it can be anticipated that the concept of corpus in CTS shall also present as one of its main attributes to allow for a wide range of configurations for data comparison.

All in all a corpus in CTS is not simply a large body of written text or spoken material as traditional definitions have often implied. It is defined more accurately as any open-ended body of machine-readable full texts analyzable automatically or semi-automatically, and sampled in a principled way in order to be maximally representative of the translation phenomenon under examination .

8.Classification Criteria for Corpora in Translation Studies

The classification criteria presented here aims to present a more flexible way of classifying the various types of corpora in translation research and pedagogy. This can be achieved by consulting [15: 229-232] typology, attempting to discuss each of the selection criteria on which corpora are generally designed along with their attributes in order to classify the types of corpora being used in the descriptive and applied branch of the translation studies as follows:

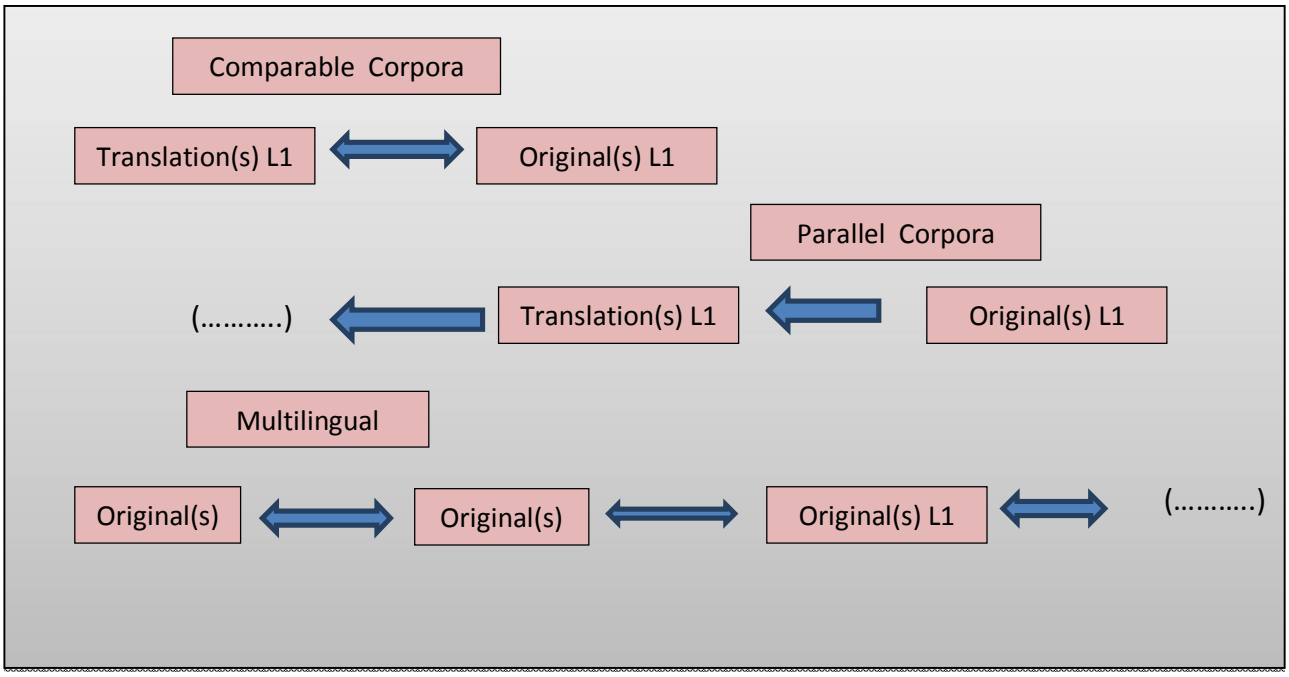

Fig. 1 – Baker's (1995) Typology for Corpora in Translation Studies

Accordingly, there are basically three main types of corpora for translation research and pedagogy:

1- **Comparable corpora**– which "consist of two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages".
2- **Parallel corpora** – consist of "original, source language-texts in language A and their translated versions in language B" .
3- **Multilingual corpora** – which are "sets of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria". Baker's tripartite classification can be re-arranged under only two main categories: comparable and parallel. This is due to the fact that the term multilingual does not have any contrastive feature that could make it distinctive from the other two types of corpora.

Translated text has always had a very raw deal in corpus linguistics. It has been specifically excluded from monolingual corpora, where it is generally treated as unrepresentative of the language being studied; irrespective of the direction of the translation .The text translated into one's native language does not normally qualify for inclusion in a monolingual corpus. Where translated texts has been studied at all, the idea has been to show that" translationese" is common, or that some of the languages that the corpus linguist is interested in, or influenced by another language. In fact, the very notion of using corpora to study translation as such –simply to

understand it as a phenomenon does not seem to have occurred to corpus linguistics. In a recent collection on corpus-based studies, [16:67] expresses the overall position of corpus linguists when she asserts that " one should refrain from using translation corpora unless the purpose of the linguistics analysis is either to evaluate the translation process or to criticize the translation product on the basis of a given translation theory".

Emphasis has shifted very gradually in recent years from the source to the target text. This did not happen overnight: for a long time, target oriented translation studies was only target-oriented in the sense of not imposing the standards of the source text onto the target text, of allowing for the fact that different contexts and communicative goals may require different translation method .

9.Processing Models and Web-based Corpora

A translation memory (TM) is essentially a database of aligned source and target texts. It can be considered a type of parallel corpus, which is interrogated in a largely automated way with the help of specialized TM software. Essentially, when a translator has a new text to translate, the TM system will automatically compare the segments (which typically correspond to sentences) contained in this new text against those stored in the databases, and if a match is found, the previous is proposed to the translator. The goal of this corpus-based technology is to allow the translator to 'recycle' relevant proportion of previously translated text. [14:23].

In computer science, the areas of parallel and distributed computing study ways in which processors can be combined in order to improve the efficiency and flexibility of a system. In computational terms, distribution can be achieved in two forms: distribution of data and distribution of processing power. One uses the expression *data stream* to refer to the flow of data from a processor to another, and analogously, *processing stream* when referring to the flow of instructions from a processor to another. There can be four classes of architectures with respect to the distribution of computational resources: SISD (single instruction, single data streams), MIMD (multiple instruction, multiple data streams), SIMD (single instruction, multiple data streams) and MISD (multiple instruction, single data streams). Although these classes are primarily used as a taxonomy of computer hardware, at least the first two of them are relevant to the way corpus software works..

10. Advantages and Limits of Computer-Corpus

According to [17:385], it becomes easy for researchers to use computer corpora classified and stored for various types of research projects. As a valid source of linguistic data, computer corpora are used in the study of syntax, phonetics and phonology, prosody, intonation, morphology, lexicology, semantics, discourse analysis, sociolinguistic variations and other areas of linguistic description. Some of the advantages of computer corpora run as follows:

1. Computers are tireless machines that can store and classify data in various ways, for instance, in studying the uses of relative clauses in academic writing and looking for their different types and function is a tiresome task if done by hand, but saving time and effort if done by computer.
2. As data are stored in an electronic format, it can be easily accessed by researchers worldwide which permits collaborative studies among various researchers.
3. Computer corpora are more practical than corpus stored in the traditional way written on sheets of paper, because they can be automatically processed and transmitted with greater speed and consistent reliability.

4. Computer corpora has strong affinity with functional linguistics as both of them claim to be closer to facts of real language use. This leaves computer corpora free from contradiction of intuition and abstraction ruling conventional methods of data collection.
5. Computer corpora are valid source of linguistic data used in various areas of language description such as, syntax, phonetics and phonology, morphology, semantics, discourse, pragmatics, sociolinguistics, language planning, etc…(see also,18).
6. The most common applications of computer corpora include: developing multilingual libraries, designing language teaching course books, compiling and developing monolingual, bilingual and multilingual dictionaries (printed or electronic) in addition to machine readable dictionaries (MRDs).
7. Functional applications of computer corpora cover the study of invariant tags which characterize meaning and usage among the varieties of a specific language, e.g., English language. The invariant tags are discourse markers like; (yeah, na, no, and eh) which often occur at the end of an utterance to provide attitudinal information .
8. The description of spoken and written register in academic and non-academic texts is among the functional uses of computer corpora. This is often carried out by examining lexical bundles distribution in these texts as lexical bundles are not only the building blocks of discourse on the phrasal and lexical levelsbut also their position effects discourse function.
9. A more subtle application of computer corpora is in building a forensic corpus to figure out the linguistic indicators of deception in police-suspects and witness interrogation .There are several linguistic cues such as, hedges, negative expressions, syntactic and semantic inconsistencies related to lexical choice and tense shifts.

Despite their numerous advantages, computer corpora have two major limitations:
1. Size plays an important role in the design of corpus because the composition of the corpus must reflect the expected goals of the research project. Thus, a corpus of investigating lexical questions should be larger than the corpus required for grammatical explorations simply because the size constraints on the latter may not affect the output as the constraints put on the former.
2. The range of language varieties used for comparing various patterns of language usage found in spoken and written discourses should encompass a wide range of possible occurrences. In this way the data derived from the corpus can be a fair representative of the possible usages across the two registers. Different registers may vary according to their being fictional-non-fictional, discourse modes, topics, socio-economic status and demographics of speakers and writers .

11. The Need for Arabic Corpus

Creating bilingual corpora containing texts from two or more languages can facilitate translation tasks, as the information available in translation can be used to create electronic data to be stored as Arabic national corpus that includes translated works from English into Arabic and vice versa. The traditional method of manual works, as carried out for the translation of Arabic perfect verb into English is manually performed on two Arabic novels written by Najeeb Mahfoudh. The researcher used a corpus of (250) sentences randomly selected from both; this is rather tiring job ,while dealing with such an issue or other similar ones in translation through computer- readable corpus or of other larger number of texts will make the task more easy and effective.[19:1-24]. Another study used a manual stylistic analysis of “Al-

Shawqiyyat" written by Ahmed Shawqi in which he cites 11,320 lines of poetry that cover (370) poems for various aspects of linguistic analysis, which is boring and time-consuming (20) . Hence, creating Arabic national corpus similar to British National Corpus (BNC), and American National Corpus(ANC) will help researchers to easily and objectively examine the linguistic features of Arabic writers in computer-readable corpora, and conduct comparisons with foreign writers.

12. Conclusions

Corpus-based translation studies have steadily grown as a disciplinary sub-category since the first studies began to appear more than twenty years ago. Although numerous studies on translation operate with corpus-based methods, most of them concentrate on the questions concerning the nature of translations and their specific features. The majority of them tried to generalize translation by defining certain rules or regularities of translated texts. Moreover, they mostly compare translations with originals, i.e., differences or similarities between translations and their source texts or comparable non-translated texts, ignoring variation which can be observed in different translation variants. Corpus-based studies dedicated to the analysis of variation phenomena involving translations, concentrate on the analysis of human translations only. However, nowadays, translations are produced not only by humans but also with machine translation (MT) systems. Furthermore, new variants of translation appear due to the interaction of both, e.g., in computer-aided translation or post-editing. In some works on machine translation the focus lies on comparing different translation variants, such as human vs. machine. However, they all serve the task of automatic (MT) system evaluation and use the human-produced translations as references or training material only. None of them provide an analysis of specific linguistically motivated features of different text types translated with different translation methods. There is also a need for Arabic corpus linguistics.

CONFLICT OF INTERESTS

There are no conflicts of interest

13. References

1.Biel, Lucja., Corpus-Based Studies of Legal Language for Translation Purposes: Methodological and Practical Potential .In Heine, Carmen & Engberg, Jan(eds.)" Reconceptualizing" LSP. Onlinepreceedings of the xv11European LSP Symposium, pp. 1-15, 2009,

2. Gries, Stefan. Th, Introduction: Corpus in cognitive Linguiatics: Corpus-based approaches to syntax. Berlin/New York: Mouton Grugter, pp. 1-7, 2006,

3.Stubbs, Michael,"Language Corpora", in Alan Davies and Catherine Elder (eds.) The Handbook of Applied Linguistics, Blackwell publishing Ltd. pp.106-128, 2004.

4.McEnery, Tony and Costas ,Garbielatos, " English Corpus Linguistics" In Aarts ,Bas and April McMahon(eds.) The Handbook of English Linguistics, Hobken(USA):Blackwell , 2006.

5.Baker, Mona, Corpus-based translation studies:The challenges that lie ahead. In Harold Somers (ed.), Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager, pp. 175–186. Amsterdam: John Benjamins, 1996.

6.Katerina Lapshinova-Koltunskl, "Variation in translation: evidencefrom corpora" In Claudio Fantinuoli & Federico Zanettin (eds.), New directionsin corpus-based translation studies, 93–114. Berlin: Language Science Press, 2015.

7.Svartvik, Jan, Corpus linguistics 25+ years on. In Facchinetti, Roberta (ed.), Corpus Linguistics 25 Years on. Language and Computers. Amsterdam: Rodopi,pp.11-25, 2007.
8.Firth, J. R., Linguistic analysis and translation. In M. Halle, H. G. Lunt. H. McLean and C.H. van Schooneveld, for Roman Jakobson :Essays on the Occasion of his sixtieth birthday 11 October, 1965.

9.Gutknecht,Christoph, "Translation" . In, Mark Aronoff and Janie Rees-Miller, (eds.)The Handbook of Linguistics Singapore: Fabulous Printers Ltd. pp. 692-703, 2006 .
10.Savory, T. H., The Art of Translation. London: cape, 1968.
11.Gellerstam, Martin , ' Translation in Swedish Novels Translated from English', in LrsWollin and Hans Lindquist(eds): Translation Studies in Scandinavian Symposium on Translation Theory:75,pp. 88-95, 1986.
12.Lindquist, Hans , English Adverbials in Translation: A Corpus Study of Swedish Renderings, Lund Studies in English 80, Lund: Lund University Press, 1989.
13.Tymoczko, Maria , ' computerized Corpora and the future of Translation Studies', Meta 43(4):pp.60-652, 1998.
14.Laviosa, Sara, Corpus-Based Translation Studies. Amesterdam-New York: Rodopi B. V. pp. 1-121 , 2002.
15.Baker, Mona, "Corpora in Translation Studies. An Overview and Suggestions for Future Research". Target, 7(2). pp. 223-243, 1995.
16.Lauridsen ,Karen, "Text Corpora and Contrastive Linguistics: which type of corpus for which type of analysis?" In Karen Aijmer *et. al*., (eds.), Lund: Lund University Press, 1996 .
17.Ibrahim, K. Riyadh, Translation oriented corpus-based contrastive linguistics .Babel 61:3,pp.381-393, 2015.
18.Biber, D, Ulla C, and Thomas A., Discourse on the Move: Using Discourse Corpus to Describe Discourse Structure. Philadelphia: Benjamins, 2007.
19.Gadallah, H.A. *et. al*., Translating Arabic Perfect Verbs into English : a text- based approach. Bulletin of the Faculty of Arts,12:1-24, 2003.
20.Al – Tarabulsi, M. A., Khasa's al-uslub fi al-Shawqiyyat[stylistic features of al-Shawqiyyat] Al-Majlis Al-Ala Lil-thaqafa, 1996.