3. Corpus and Translation Studies

3.1. Introduction

Corpus linguistics has played a vital role in a variety of researches in the field of applied linguistics while also contributing to researches dealing with languages including journalism etc. In its earlier days the corpus was being stored in large mainframe computers which involved punching or storing the data on tapes; the constraints were mainly due to the processing capabilities of the computers as well as the storage technologies of those times. Further, there were problems encoding characters with accents. With the developments and evolution of technology, i.e. the massive improvement in the processing capabilities of the computers, as well as the newer storage technologies, the applicability of corpora in researches has been drawing the attention of the researchers who view corpora as a tool that could provide insights into various aspects of language and linguistics. The technology has enabled the corpora to be used for conventional researches not only in the field of linguistics but also in other related fields like translation studies. Further, corpus studies which was more of a tool for researches has now become an object of study in its own right, exploring possibilities of extension to other fields of research.

According to Biber (Biber 1998) corpus linguistics is a method which is characterized by four traits:

- It is empirical; analyzes the actual patterns of use in natural texts
- It utilizes large and principled collection of texts, known as “corpus” as the basis of analysis
- It makes extensive use of computers for analysis, using both automatic and interactive techniques
It depends on both quantitative and qualitative analytical techniques

3.2. What is Corpus?

Corpus can be defined as a collection of knowledge on a certain topic. In the field of linguistics varying definitions of corpus have been put forth by scholars. These definitions could broadly be classified into two categories. While the first type defines corpus in a very rigid framework, the second type takes a very liberal and flexible approach towards the term corpus. Gabriela Saldanha says:

Because there is no unanimous agreement on the necessary and sufficient conditions for a collection of texts to be a corpus, the term ‘corpus’ can be seen in the literature referring sometimes to a couple of short stories stored in electronic form and sometimes to the whole world wide web. In order to discuss the fundamental principles of corpus linguistics, it is important to first establish certain limits around what can and cannot be considered a ‘corpus-based’ study of translation

(Saldanha, 2009:07)

She cites following definitions set forth by various researches in support of these observations:

McEnery and Wilson (1996: 87), emphasize representativeness: “a body of text which is carefully sampled to be maximally representative of a language or language variety.

According to Bowker and Pearson (2002: 9) a corpus is: “a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria”
Leech (1992: 106) views corpus as: “a helluva lot of text, stored on a computer”.

For Kilgarriff and Grefenstette (2003: 334): A corpus is a collection of texts that is considered as an object of language or literary study.

Although all the above definitions commonly mention a collection of texts, there are differences in terms of size and nature of the corpus being talked about. While McEnery and Wilson’s definition talks about body of text that is carefully sampled so that it is “representative” of certain traits of the language. Saldahna comments that this is a rigid approach towards the corpus as it is difficult to define what exactly the term “representative” would mean, therefore, making it a difficult criteria for corpus selection. On the other hand, other definitions provide flexibility in their approach towards corpus by mentioning “collection of texts” with large “size” of corpus. However the term “size” in the all the above definitions differ.

By and large, the term “corpus” can now be summarized as any collection of texts which have been stored systematically in electronic and computerized formats.

3.2.1. Corpus in Language Research

Corpus linguistics is a method that deals with the study of language through the analysis of corpus or data that is held in electronic format.

The analysis of data is carried out by employing software and tools specially developed to retrieve the relevant information from the corpus (also known as mining) according to the criteria and parameters provided by the information seeker. The researchers in turn subject the retrieved data to various analyses for finding relevant answers to their research questions. Such analyses play a vital role in developing theories related to linguistic and language use. The analysis
could be on different levels of language such as lexical, semantic, grammar, culture etc. The data analyzed is from a collection of data or texts that are authentic or attested i.e. to say that these cases are naturally occurring language instances (it is not the kind of data that is created for the purpose of a specific research).

3.2.2. Types of Corpora

- General corpora - Corpora containing a large collection of diverse forms general in nature such as written texts and spoken language used by speakers of different regions, genders, ages, social classes (could be economic, sociological, professional etc.) that have occurred naturally in general settings. Examples of such corpora are Collins Wordbanks Online English Corpus, British National Corpus or the Bank of English.

- Synchronic corpora- This kind of corpora consists of data pertaining to language in a particular point of time in history enabling the researchers to look at various elements of language at that particular point of time; for example F-LOB and Frown Corpora.

- Historical corpora or Diachronic corpora - Contains texts from various eras and periods in History. Such corpus allows researchers to study various linguistic aspects diachronically. For example change in language use spanning across periods of time. Examples of such corpora are ARCHER and the Helsinki corpora.

- Learner corpora – Contains collections of representative texts that has been produced by the language learners themselves. Such corpus is useful in facilitating foreign language learning. For example International Corpus of Learner English and the Cambridge Learners Corpus.
- Regional corpora – Contains collection of instances of the same language used in different regions; such corpora focuses on the representative traits of a certain language in a particular region. For example Wellington Corpus of Written New Zealand English.

Multilingual Corpora – Contains written texts or spoken forms related to two or more different languages. Such corpora find use in contrastive or comparative studies between different languages. (online quoted by Najda Nasselhauf, and by Friederike Müller and Birgit Waibel)

3.2.3. Corpus Tools

Following are some examples of tools available related to Corpus:

**Multi-Lingual Parallel Concordancer**

This a concordancing program for parallel texts developed by David Woolls and others in the Lingua project.

Platform: MS Windows.

**ParaConc**

ParaConc is a tool designed for linguists and other researchers who wish to work with translated texts in order to carry out contrastive language studies or to investigate the translation process itself.

Platform: Mac.
WordSmith
Software developed by Mike Scott and published by Oxford University Press. It can perform lexical analysis of texts and alignment of multi-lingual texts.
Platform: MS Windows 3.1 or higher.

3.3. Role of Corpus in Linguistic Research
3.3.1. Types of Corpus Research

There is a need to differentiate between terms Corpus research, Corpus based research and Corpus driven research.

Corpus Research involves researches that primarily relate themselves to the design of the corpus. Corpus methods, tools used for mining or extrapolating data, models related to annotating, abstracting and analyzing etc., anything that might become a corpus framework enabling the researchers to carry out corpus based or corpus driven studies. In short corpus research directly focuses on the corpus as a field of research.

Corpus based research identifies itself with the researches that attempt to validate the theories proposed by scholars based on their researches. Biber (Biber 2004; 2009) in his article says that “the goal is not to discover new linguistic features but rather to discover the systematic patterns of use that govern the linguistic features recognized by the standard linguistic theory” Biber (2010:167). He says that the contribution of the corpus based approach is that the findings “often” contradict the prior “intuitions” of the linguists In support of his comment he cites two case studies where the corpus based study has proven that the predictions were based on faulty intuitions.
Corpus driven research on the other hand involves finding and proposing new theories that were not existing previously while using the corpus for finding new evidence identifying new linguistic categories or/and units. Similar distinction has been made by Tognini-Bonelli (2001) which has been quoted in the articles by Biber as well as Saldanah.

Biber notes that strictest form of corpus driven research assumes only the existence of word forms. He distinguishes between these studies while noting that the researchers who have used the corpus driven research for investigating the formulaic aspects of language placing their studies in the purview of “lexical bundles” by looking at simple word forms for the frequency of occurrences to identify recurrences. He argues that there is a difference between the “lexical bundles” and “formulaic language”. He says that lexical bundles are multi-word sequences that recur frequently and find wide distribution across different types of texts. However, they are not necessarily complete structural units recognized by current linguistic theories. Therefore they differ from the formulaic language in three major ways:

- They are extremely common
- They do not possess idiomatic meaning
- They do not represent complete structural units.

In contrast to this, current linguistic theory recognizes formulaic language to be a complete structural unit and idiomatic in nature. He says that the findings of the corpus analysis indicate that this popular opinion does not hold good and formulaic expressions possessing such characteristics are extremely rare.
3.3.2. **Broad Fields of Research in which Corpus is Used**

The term ‘Corpus Linguistics’ does not identify itself with any one particular field of research but is associated with variety of fields covering various heterogeneous areas ranging lexicography, translation studies, descriptive linguistics, applied linguistics etc. and also fields where introspective methods are unable to answer questions such as studies of language variation, dialect, register and style, or diachronic studies (Jacqueline Léon, 2005). Although there are diverse fields that attempt to exploit the potential of the corpus for researches in their respective researches, all of these fields share the use of the corpus of various kind of texts in written or spoken forms for extrapolating information relevant to the research question. The investigation is more of data-driven rather than a rule-driven analysis.

3.3.3. **Research Questions Corpus can Answer in Researches**

Gena R. Bennett in her book (Bennett Gena R, 2010:4 ) highlights what corpus linguistics can do and what it cannot. She begins by highlighting the kind of questions it attempts to answers:

- What are the most frequent words and phrases in English?
- What are the differences between spoken and written English?
- What tenses do people use most frequently?
- What prepositions follow particular verbs?
- How do people use *can, may*, and *might*?
- Which words are used in more formal situations and which one are used in more informal situations?
- How often do people use idiomatic expressions?
- How many words must a learner know to participate in everyday conversation?
• How many different words do native speakers generally use in conversation? (McCarthy 2004 quoted by Bennett, 2010:4)

Later she also adds what Corpus Linguistics cannot do:
• Providing negative evidence
• Explaining why?
• Providing all possible language at one time? (McCarthy, 2004 quoted by Bennett, 2010)

According to a survey by Gilquin & Gries (2009:3) corpus-linguistic studies published over the course of four years in three major corpus-linguistic journals were mostly:

• exploratory (as opposed to hypothesis-testing) in nature;
• on matters of lexis and phraseology, followed by syntax;
• based on written data;
• using frequency data and concordances, followed by simple association measures.

3.3.4. Methods in corpus Linguistics
This part of the thesis attempts to look at different methods used for exploiting the corpora. The data collected in any form is raw data and may not be useful unless it is processed such that it fulfills the criterion that defines “Corpus”. As it has been mentioned earlier also, corpus is a collection of texts systematically stored in an electronic format and can be retrieved according to the parameters entered by the researcher as per his or her requirements using software tools like Wordsmith etc.

There are three types of information that are contained in the corpora that make the extrapolation of the data easier (McEnery, Hardie).
These are:

**Metadata** i.e. the information about the data itself e.g. author, publication time, and the language so on and so forth. Metadata can be embedded in the corpus or alternatively be held independently as document or database. There are diverse kinds of corpora available, some that fit into the conventional confines of the term “Corpora” and some that are deviant forms of corpus, for example, audio and video recordings (Gries, Berez online), there are equally diverse forms of annotations in use.

**Textual Markup**, i.e., the information other than the textual elements such as the textual format etc. This kind of information is useful for transcript analyses.

**Annotation** i.e. tagging various linguistic elements of the texts with their respective traits.

Annotation is the first step to make the data retrievable. The data is stored in Unicode character format to remove the dependencies attributed to different character sets of different languages in other words making the characters recognizable universally by any system. The next step is to annotate (also known as tagging) the data using eXtensible Markup Language (XML). Gries and Berez list out some types of annotations:

- Lemmas
- Part-of-speech tagging: syntactic and morphological annotation
- Syntactic parse trees
- Semantic annotation
- Phonetic and phonological annotation
- Prosodic annotation
- Sign language and gesture annotation
- Interactional annotation
3.4. Corpus and Translation Studies

The commentary on the role that corpus can play in the researches pertaining to Translation Studies begins with Mona Baker who attempts to introduce concepts of corpus into translation studies. Baker observes that translation studies have heavily relied on introspective methods for data collection and that scholars want to look at large number of instances as means of investigation in Translation Studies. Baker cites Holmes and Newman voicing out their dissatisfaction overreliance on introspective methods. Baker quotes Holmes (Holmes 1988, as quoted by Baker, 1993):

Many of the weaknesses and naiveties of contemporary translation theories a result of the fact that the theories were, by and large, developed deductively, without recourse to actual translated texts-in-function, or at the best to a very restricted corpus introduced for illustration rather than verification or falsification.

(Holmes 1988:101, as quoted by Baker, 1993:240)

Newman similarly suggests that for finding generalities for the purpose of theorizing there is a need to examine actual instances to determine these generalities.

Baker (1993) notes that translation studies has not been given much attention in corpus linguistics as the translated texts have been constantly been excluded from the monolingual corpus assuming that not much can be studied about the language in question. Corpus linguists have argued that the use of translation corpora should be limited only to studies that attempt to investigate the translation process or to study a particular translation as a product based on a specific theory of translation (Lauridsen 1996: 67).
Traditionally, ever since the descriptive translation studies has come in to being, only one kind of corpus has been recognized i.e. parallel corpus where the source text is available along with its translation. According to Baker this approach of imposing limitations might curtail the potential of corpus in empirical research of translation as a process or as a product. Baker observes that all kinds of corpora may be admitted without any assumptions pertaining to the value that the corpus might hold with regard to the results that a study might generate. Baker suggests that a different kind of approach needs to be adopted while looking at corpus that may be used to look at language of translation, without judging on the basis of preconceived notions.

3.4.1. Corpus in Descriptive Translation Studies

In Descriptive Translation Studies corpus has been used to carry out contrastive and comparative studies such as how different or similar a translation is from its original text or comparing the different translations of the same text to study the stylistic and cultural factors that affect the translator. Scholars in the field of Descriptive Translation Studies have voiced the need to bring together the fields of Descriptive Translation Studies and corpus linguistics. The reason being corpus linguistics is a method that is used to observe and identify recurrent patterns in a language statistically. Therefore, it may be assumed that it may prove to be a reliable method to identify generalities and theorize phenomena related to a language. Since translation is an activity that involves language, corpus methods may prove to be useful in Descriptive Translation Studies as well.

There are two kinds of applications of corpora in Descriptive Translation Studies. The first application deals researches in Translation Studies e.g. using corpora to validate the existing theories or identifying general phenomena in translations and propound theories with regard to these. The second one deals with using corpora as a resource for reference by translators and for training translators, and also for
searching terms in contexts in cases of advanced or professional translators. According to Mona Baker there are three kinds of corpora that may be applicable for researches in DTS:

- Parallel Corpora
- Comparable Corpora
- Multilingual Corpora

However, the question still remains whether these existing corpora are suited for every kind of translation research. Most of the corpora available today were or are primarily developed for catering to linguistic researches, at the most providing instances for the purpose of machine translations or learner and professional translators who often use these as a resource for performing their translational tasks.

### 3.4.2. Available Parallel, Multilingual, Comparable Corpora

#### 3.4.2.1. English-German Translation Corpus

The English-German translation corpus is a collection of texts from various English and German sources. The texts are collected from diverse fields ranging from literature to scientific texts. This aims at creating a corpus of texts that are aligned and are machine readable that would provide scholars a means to carry out empirical investigation to discover and categorize translation equivalents for various linguistic elements such as prepositions, function verbs, deictic elements, metaphors or culture-specific structures. Based on the patterns of occurrences that emerge from the corpus search results, affective factors can be generalized; for example the common choices of the translators while translating.

The corpus provides a means to provide insights into contrastive and cognitive aspects of translation process. The findings of such studies can
also be extended to other related fields such as bilingual lexicography, language learning/teaching, improving translation aids etc.

According to this corpus project website, the analysis of the corpus is being undertaken while partnering with scholars and institutions involved in similar kind of projects in order to stimulate inputs into fields such as translation studies, corpus linguistics etc. and analyze closely related languages. This project particularly focuses on two parallel projects analyzing English-Swedish (Lund) and English-Norwegian (Oslo/Bergen) language structures, so that English is studied in contrast to three Germanic languages in the same methodological framework.

3.4.2.2. English-Norwegian Parallel Corpus (ENPC)

The parallel corpus is an open collection of texts that have been compiled as a general research tool intended for researchers engaged in applied and theoretical linguistic research. The corpus is being compiled as a part of a joint project undertaken by the Department of British and American Studies, University of Oslo, and the Norwegian Computing Centre for the Humanities, University of Bergen. The project also focuses on the computer processing of parallel texts. The coding system used to mark up the ENPC follows the suggestions made by the Text Encoding Initiative (TEI) as presented in Guidelines for Electronic Text Encoding and Interchange (Sperberg-McQueen & Burnard, 1994). The corpus uses Start- and end-tags (<..> and </..>) as a markup for various elements of the texts. The most important tags are paragraph markup (<p>...</p>) and sentence markup (<s>...</s>). They have been, coded, and proofread after which they are aligned at sentence level, i.e. each original sentence is linked to the respective translated sentence. The alignment of the sentence is carried out automatically using a program developed by Knut Hofland,
subsequently the text is proofread manually. The texts stored in the
database can be searched through Translation Corpus Explorer, which is a
browser developed by Jarle Ebeling.

The compilation of the English-Norwegian corpus was completed in 1997.
More languages were added (German, Dutch, Portuguese) between 1997-
2001. The tagging of the parts of speech has been completed for the
English and the Norwegian original texts.

The focus of the corpus has been on novels and general non-fictional
books. The corpus contains chunks extracted from the texts from various
authors and translators. The limit was imposed so that more variety and
styles could be included in the corpus. The corpus contains 30 original text
extracts in each language and their respective translations, whereas the
non-fiction part contains 20 texts in each direction.

The project aims (as cited under aims on the project website) to:

1. Compile a parallel corpus of English and Norwegian texts for
   computer processing
2. Develop tools for analyzing parallel texts
3. Carry out studies of the structure and communicative use of the
two languages on the basis of the corpus.

Areas for study include:

- representative constructions in English and Norwegian (Jarle
  Ebeling)
- word order and information structure in English and Norwegian
  (Hilde Hasselgård)
- lexical comparison of English and Norwegian (Kay Wikberg)
Examples of more general questions to be studied are:

- To what extent are there parallel differences in text genres across languages?
- In what respects do translated texts differ from comparable original texts in the same language?
- Are there any features in common among translated texts in different languages (and, if so, what are these features)?

The studies attempt to investigate the general and language specific phenomena. According to the scholars involved in the project, the studies and their findings are therefore useful from the point of view of understanding the language itself in general, as well as, for the studying the respective languages that are being compared. The findings of the researches find application in the fields of lexicography, language learning and teaching, and translation studies.

3.4.2.3. English-Swedish Parallel Corpus (ESPC)

The English-Swedish Parallel Corpus (ESPC) consists of a collection of fictional and non-fictional original texts and their translations (English to Swedish and Swedish to English). The original English and Swedish texts are comparable in terms of size and text type. The total size of the corpus is 2.8 million words. The parts of speech tagging has been completed recently. Corpus is compiled on the lines of the English-Norwegian Parallel corpus.

The present version of the ESPC contains 64 English texts and their translations into Swedish and 72 Swedish texts and their translations into English. There are more fictional texts as compared to non-fictional texts. However, they are almost same in terms of corpus size. The corpus
contains complete works in cases of smaller texts and chunks of text ranging from 10000-15000 words extracted from larger texts. The characteristic feature of the structural design of ESPC is that it can be used as a parallel corpus as well as comparable corpus in following permutations and combinations:

As a parallel corpus-
- Comparing the English original text and its Swedish translation
- Similarly, Swedish original text and English translation.

As a comparable corpus-
- Comparing original Swedish text and English text
- Comparing the original texts and the texts translated into the same language
- Comparing translated texts in different languages

3.4.2.4. The Intersect Parallel Corpus (English-French)
The INTERSECT is a parallel corpus compiled by Raphael Salkie et.al at the University of Brighton. The corpus includes English texts and respective translated versions in either French or German stored in an electronic format. The texts are authentic and have been collected from diverse genres such as fiction, non-fiction, mass media and communication, business reports, UN and EU documents, science and technology, tourist literature etc. The size of the corpus is about 1.5 million words in French-English pair and about 800,000 words in German-English pair.

Each source text has been carefully analyzed and translated into the target language by experienced translator. The corpus provides word equivalents
beyond the words included in conventional dictionaries according to occurrences in their natural contexts. The corpus is stored in a plain text format and are aligned using a parallel concordance praracone developed by Michael Barlow at the University of Auckland, New Zealand. The project members are presently working on further annotating the linguistic elements of the texts. The corpus has been used for lexicographical research at the University.

The texts contained in the corpus according to the website include text of following nature.

French-English corpus:

- Articles from Le Monde and their translation in Guardian Weekly
- Magazine articles and official documents from Canada
- Instructions for a variety of domestic appliances
- Technical texts about telecommunications
- Texts from international organisations
- Modern fiction
- Academic textbooks

3.4.2.5. The German-English corpus

- Various company home pages: Hoechst, Siemens, BASF
- The constitutions of Germany, Switzerland and Austria
- Various EU documents
- Transcripts of speeches by the President of Germany
- Short news items from the 'German news' web site
- Various United Nations documents
3.4.2.6. Corpus Resources and Terminology Extraction (CRATER) Corpus

Corpus Resources and Terminology Extraction (CRATER) is a corpus that has been compiled as a part of a joint project between Lancaster University, England and Universidad Autónoma de Madrid, Spain. The entire corpus consists of technical texts from the International Telecommunications Union (ITU) written in English, French and Spanish. The corpus consists of 5.5 million words. The corpus is aligned at various levels ranging from sentences to word using a number of specialized alignment algorithms. Further, the texts are manually tagged with part-of-speech and morphological annotation.

The Multilingual Parallel Corpus (Danish, English, French, German, Greek, Italian, Finnish, Portuguese, Spanish, Swedish texts)

3.4.2.7. Portuguese – English Parallel Corpus

Portuguese – English Parallel Corpus, an open-ended corpus collection of Portuguese – English and English – Portuguese sources texts and also machine searchable corpora, is developed on the core structure of the English- Norwegian Parallel corpus (Johansson et al. 1999). The machine readable and searchable collections of source texts are originally written in Portuguese and in English that have been aligned with their respective language translations. As of now, Compara contains only published fiction texts, however, other genres are expected to be added in upcoming stages. Compara, which has been developed under the broader framework of the Computational Processing of Portuguese project, has been conceived not only for users experienced corpus user but also for users who not necessarily Corpus –literate.
Text Selection:
1. Academic Prose
2. Journalistic Text
3. Instruction Booklets
4. Tourist Brochures
5. Fiction

Alignment

The Aligner used in Compara is an Easy Align tool (v.1.0) written for use with the IMS Corpus workbench. This tool does not single out source text sentence divisions as being more important than translation text. Whenever source and translation sentence division do not correspond the tool either matches with many other sentences, which is permissible according to Compara’s alignment or it matches many source texts to the translation text.

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>S</td>
<td>S,S</td>
</tr>
<tr>
<td>S</td>
<td>½S</td>
</tr>
<tr>
<td>S</td>
<td>null</td>
</tr>
<tr>
<td>S</td>
<td>S(+S)</td>
</tr>
</tbody>
</table>

Fig.3.1. Alignment image

Compara, first came into use in January 2001, and after couple of years from the day it came into use, Compara had permission to include extracts of 60 different Portuguese-English text pairs by authors and translators from Angola, Brazil, Mozambique, Portugal, South Africa, United
Kingdom and United States. These texts are combined product of the work done by 33 authors and 31 translators. Approximately, COMPORA has total of 190,000 words in each language

3.4.2.8. **The Scania Corpus**

Scania Corpus is a corpus consisting of a large collection of truck manuals of major Swedish company Scania CV AB. The Department of Linguists of the Uppsala University (UU) brought this corpus into use in cooperation with Scania for developing translation support tools. Initially, this corpus was delivered to UOU in Frame maker Format. However, later is has been converted to TEI-Compliant SGML. The corpus contains documents in eight European languages namely, Swedish, English, French, Dutch, Spanish, Finnish, German and Italian, among which Swedish is the source language. The total size of the corpus is over 2 million words. In the year 1998 documents (mainly Swedish and English) were converted to Uppsala’s SGML format using TEI lite.

Alignment: The Corpus uses GC-Align Program for aligning the sentences, of translated documents, in eight language versions.

3.4.2.9. **The SCARRIE Swedish Newspaper Corpus**

The SCARRIE Swedish Corpus was developed by the Department of Linguistics at the Uppsala University. The project was based on the input of words from two Swedish language newspapers – Svenska Dagbladet and Uppsala Nya Tidning. The project started in November 1996 and has been available since June 1999. It was an EU-funded collaborative project to develop a high-quality proof-reading tool in Scandinavian languages. Although the project is closed now, the availability was restricted when it ran, some of the deliverables are available to a larger community. The
corpus consists of about 70 million words including single letters, bigrams and trigrams.

3.4.2.10. Swedish Political Texts
This corpus contains texts from the Swedish government among that the declarations issued by the Swedish prime-minister once a cabinet starts afresh. These declarations are issued in 5 languages: Swedish, English, French, German and Spanish (since 1996). The present size of the corpus is 11,000 words.

3.4.2.11. A multilingual term bank of car maintenance terms
Another Offset of a Scania Corpus brought by The Department of Linguists of the Uppsala University (UU). This data bank consists of approximately 4000 Swedish car maintenance terms with translations in 6 other languages: Dutch, English, French, German, Italian and Spanish.

3.4.2.12. ACL SIGLEX Parallel Corpora
The ACL SIGLEX Parallel Corpus, a bilingual parallel corpus of English and French, containing EC official documents on telecommunications. The corpus is a part-of-speech tagged and additionally legitimatized. Currently the 2 sub-corpora are held one after the other, but will soon be loaded right into a database bundle (INGRES) permitting retrieval of parallel bilingual text (McEnery and Daille, 1993). It consists of about 1,250,000 words from the respective languages.

3.4.2.13. BAF Corpus
The BAF is a French – English corpus, which is produced by researchers at the CITI, a Canadian government research laboratory, as part of their contribution to the Action de recherché concerteé (ARC), a project initiated and financed by the
AUPELF-UREF. BAF consists of pairs of English and French documents, which are translation of one another and whose sentences have been aligned. While, lingual corpora are now quite common, the alignment which is entirely done by hand makes this BAF from others. Hence, the corpus can be used as a reference to evaluate and compare the performances of various automatic techniques. As for the contents are concerned, BAF is made up of “institutional” texts; debates that took place in Canadian parliament, Legal transcripts and UN reports, which includes both technical and scientific documents. In all, the corpus contains approximately 400,000 words in each language. The Alignment of the BAF is finer and has record correspondences at the level of sentences. This decision was based on a number of factors, such as, sentence level alignments, word level alignment.

The Protocol Alignment is as follows:

1. Segment boundaries always coincide with sentence boundaries.
2. The nth segment in one text and the nth segment in the other are translations of one another.
3. Segments are always as small as possible.

File Formats:

1. COAL format: This is the format in which the alignments were originally produced. A pair of documents in COAL format consists of three distinct files: two plain text files, and an alignment file. The alignment file contains a sequence of pairs \([(s_1, t_1), (s_2, t_2), (s_n, t_N)]\) where each pair corresponds to a segmentation point, expressed as a pair of character offsets.

2. CES format: In this format, three files also correspond to each pair of documents: two text files (CESANA format) and an alignment file
(CESALIGN format). The text files are enriched with SGML mark-up that uniquely identifies each sentence in the text.

The only difference between the COAL and CES formats is that the CES assumes a complete segmentation of the texts into sentences, which is made explicit by the mark-up. The COAL format does not make that segmentation explicit, and nothing guarantees that the segmentation that is implicit in the alignment is complete. For example, if one sentence in the English text is translated as two sentences in the French text, the boundary between the first and second French sentences will not appear in the COAL alignment.

3. HTML format: This is a "visualization" format. Here, to a pair of documents corresponds a single HTML file, which can be loaded into any HTML viewer capable of displaying tables and colours.

Fig.3.2.BAF parallel corpus retrieval image
Source: BAF Website

3.4.2.14. CRATER Multilingual Aligned Annotated Corpus

The CRATER corpus, which is developed on the foundations of a previous project, named ET10/63, a project funded in the final phase of the Eurotra programme. This is this produced by the Department of
Linguistics & Modern English language of The Lancaster University in United Kingdom in getting, with the financial help of ELRA. The Corpus Resources and Terminology extraction project (MLAP-93-20) has extended the bilingual annotated English-French International Telecommunication Union corpus produced with ET10/63 to include Spanish. This version being called CARTER-2, is expanded version of French, English component of the parallel corpus by increasing the size of the English and French corpus from 1,000,000 words per language to approximately 1,500,000 words per language. The offer consists of 1,500,000 tokens for English and French and 1,000,000 token for Spanish, with manually edited morpho syntactical annotations.

3.4.2.15. **Multilingual Text Tools and Corpora (MULTEXT)**

Multext envelops a progression of ventures whose objectives are to create models and details for the encoding and processing of linguistic corpora, and to create tools, corpora and linguistic resources encapsulating these benchmarks. Multext is developing tool, corpora, and linguistic resource for a wide assortment of languages, including Bambara, Bulgarian, Catalan, Czech, Dutch, English, Estonian, French, German, Hungarian, Italian, Kikongo, Occitan, Romanian, Slovenian, Spanish, Swedish and Swahili. All Multext results are made public and freely accessible for non-business, non-military purposes.

3.4.2.16. **MULTEXT-East**

The MULTEXT-East assets are a multilingual dataset for language engineering research and development. This dataset contains, for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Macedonian, Persian, Shine, Resian, Romanian, Russian, Serbian, Slovak,
Slovene, and Ukrainian, some, or the greater part of the accompanying language resources: the MULTEXT-East morphosyntactic details, lexica, and annotated "1984" corpus; the MULTEXT-East parallel and comparable and speech corpora; and related documentation.

3.4.2.17. Tran learn

Corpus consists of parallel texts in English and French, drawn from official records of the proceedings of the Canadian Parliament. Whereas the content is restricted to legislative discourse, it spans a broad assortment of topics and also the rhetorical vary includes spontaneous discussion and written correspondence at the side of legislative propositions and ready speeches. The collection bestowed here has been assembled by the LDC by means of archives from 2 distinct secondary sources. Material from one time period of parliamentary proceedings was acquired through the IBM T. J. Watson Research Centre, while material from another period was acquired through Bell Communications Research Inc. (Bellcore). The combined assortment covers a time span from the mid-1970 through 1988, with no apparent duplication between the 2 information sources.

Common features between the three sets:

- They are rendered here using the 8-bit ISO-Latin1 character encoding standard.
- They use a minimal amount of SGML tagging to identify sentences or paragraphs.
- All sets are organized using a parallel file structure, in which the content of a given English text file is matched by the content of a corresponding French text file.
The SGML text files for the IBM and the Bellcore committee-hearings data are published in compressed form, using the public-domain GNU-Zip utility (gzip). The Bellcore main-session files are not compressed.

In terms of differences between the three sets:

- The IBM collection is presented as a sequence of parallel sentences (there are nearly 2.87 million parallel sentence pairs in the set).
- The Bellcore data are presented as sequences of paragraphs.
- The Bellcore main-session data is accompanied by mapping files that provide computed paragraph alignments and word-token correspondences; no additional alignment data are provided for the Bellcore committee texts (and none are needed for the IBM sentences).

### Detailed information was not available for the following Corpora

1. The Swedish Immigrant Newspaper corpus
2. Bilingual Reference Corpora for translators and translation studies
3. English Turkish Aligned Parallel Corpora
4. FECCS
5. Knowledge Acquisition for Japanese-English Machine Translation
6. Michel Barlow’s Parallel Corpora
7. PEDANT Parallel Texts in Göteborg

After studying at the existing parallel corpora available, it emerges that the existing parallel corpora contains only text in source language and translated text into targeted language. The researchers engaged in process research seek a completely different kind of information as we have seen in the Review of Literature. Therefore, most researchers have created their own data by conducting
experiments to study the process of translation. However, the information and data elicited through these experiments has not been put in a public domain so that other scholars can refer to the data for their own research. According to Susanne Gopferich, if the data can be made available it would help researches in Process Studies in the following in two ways:

1. The results may be verified by other researchers by attempting to reproduce the results.

2. The researchers can use the same data from the corpus for their respective researches.

In her paper “Data Documentation and Data Accessibility in Translation Process Research”, she suggests that the corpus for Process Oriented Studies should:

- contain Meta-data (describing all data elicitation methods used, the description of subjects and the researchers, nature of study etc.) that would help the researchers not only in interpreting the findings based on the corpora but also help them decide whether the data is suitable for his or her research
- be accessible via internet so that it can be reused by researchers
- be retrievable matching the search query with the information stored in the meta-data
- contain data archived in such a way that the specific type of data can be retrieved by applying search criteria (e.g., data resulting from one specific project or data from all experiments that match the search criteria)
- contain transcripts of audio recordings (preferably in a machine readable format e.g. XML generally used for annotating documents)
- include video recordings
3.5. Conclusions

In this chapter, the corpus methods and procedures adopted in conventional linguistic research were studied following which the corpus methods with reference to DTS were examined. The objective was to understand how and to what extent corpus methods are being applied in Translational Researches and further to find if there are any existing corpora conducive to Translation Process Research. It was found that parallel corpora i.e. corpus of aligned source and translated texts exist. The corpora ranged from general fictional and non-fictional texts to texts pertaining to specialized fields. The websites were also referred to for understanding different aspects such as size, nature of researches aims that the corpus intends to fulfill. General aims stated were:

- Lexicographical use
- Use in Contrastive studies
- Translator training
- Resource for the reference of professional translator
- Investigating the Translation Process

The kind of information that the process research seeks is a different kind of information that is elicited using a combination of multiple methods. While the conventional parallel corpus focuses on the equivalence of words and other linguistic elements like metaphors etc., the data created by the process researchers focuses on information relating to the problems faced, the solutions provided, the resources used and the strategies adopted by the translators.

From the study it emerges that the existing parallel corpora may not be suitable for Process Oriented Study of Translation as the data included is only in the form of the source text and the target text, whereas the process researchers prefer mixed
methods approach that facilitate triangulation, in addition to the source text and target text as a basis for theorization.

The present chapter highlights two main points:

1. Adoption of Triangulation methods for researching the translation process.
2. Use of technology to facilitate triangulation.

If the above two points can be integrated with conventional corpus methods, researchers may be able to use the data contained in such a corpus resulting in saving time and effort involved in eliciting their own data.

Bearing this in mind, in the next chapter I would first like to attempt to understand what triangulation is and how it can contribute to quantitative and qualitative research. Following which I would like to evaluate software that can be integrated with the corpus model that could facilitate triangulation in Process Research of translation.