Translating the Language of Aviation. The Development and Detailed analysis of the English-Bengali Aviation Corpus for Machine translation

Saptarshi Paul

Department of Computer Science
Assam University
Silchar, Assam, India
paulsaptarshi@yahoo.co.in

Abstract. The recent advent of corpora based transliteration and translation approaches such as SMT and NMT models are completely based on the parallel corpus. It is the corpus that ultimately decides the Translation Accuracy (TA) of the model. With the regular and common domains exhausted and things of the past, Modern fields of research corpora domains lies anywhere between medicines to aero-science. The Work becomes more interesting when Indian languages are taken up specially ones that include technical touch such as Aeronautics and Aviation. With corpora for technical domains in English-Indian languages pairs such as Bengali coming up now, the automatic analysis for such corpora are an interesting aspect that researchers are taking up. Such analysis also helps developers and researchers to further improve the quality of the corpus and set new benchmarks for development of future corpora. This paper deals with the need, development and detail analysis of a bilingual corpus in aviation for English and Bengali language pair.

Keywords: Aero-science; parallel-corpus; text-analyzer; aviation

(Received July 19th, 2022 / Accepted September 1st, 2022)

1 Introduction

Recent trends of Translations and Transliteration work are corpora dependent. Corpora are considered as one of the most important linguistic tools. The quality, fluency and adequacy of the corpus directly affect the percentage of translation accuracy of the model (NMT or SMT) [17, 16]. That is, we can say models are crafted using the parallel-corpora. NMT systems are dependent on bilingual parallel corpora. Parallel corpora are used to train NMT and SMT systems. The composition and size of a corpus determines the accuracy and efficiency of an NMT system. Bilingual parallel corpus is a collection of texts in one language (Source) and their translation into other language. For our research we consider English and Bengali language. The text of source language is translated into the text of target language and both are aligned, i.e. both needs to be matched. Parallel corpus is considered one of the primary linguistic resources in NLP. They are used as an integral part of many NLP tools some of which are summarized below:

1. E-dictionary: With parallel corpus we can create bilingual or multilingual E-dictionary. Here we have created a bilingual E-dictionary for the aviation domain to assist in the creation of the English-Bengali aviation corpus.
2. Machine Translation: Parallel corpora are an integral part of any modern machine translation system. Both SMT and NMT use parallel

corpora.

3. Information Retrieval and Extraction: Natural language information such as grammar, phonemes etc can be retrieved from parallel corpora.

4. Predictive analysis: Nowadays predictive analysis is found in all spheres of research, academic and practical applications. Aircraft maintenance, medical applications, legal suggestions, travel guidelines and agricultural practices, all use predictive analysis. Parallel-corpuses are composed of parallel sentences in the source and target languages. Researchers are involved in the development of perfect corpuses and as well in developing various tools that can assist in both the creation and analysis of the parallel corpus. This work is an off-shoot of a bigger project for the creation of an English-Bengali Aviation Translation system [15]. While building the said NMT system the need was felt to develop an analyzer tool that could perform word count, concordance and other analysis on the English-Bengali corpus created specifically for the project. This tool would later go into playing a lead role in not only analyzing the parallel corpus but also perfecting it. As such a corpus was being developed for the first time so it was to build the benchmark for future works. While creating the corpus, certain management aspects were kept in mind, like using as many aviation OOV words and phraseologies as possible to use them from different perspective. The Paper follows a systematic pattern that can be described as follows: the second part of the paper describes the related research work done in the field of both NLP applications and corpus creation. The third part describes why we need the aviation corpus, the analyzer tool and the unique vocabulary composition. The fourth and fifth part of the paper describes the composition of the corpus and how it was created. The details of the creation of the analyzer tool and its output are described in the Sixth and seventh part, while analysis has been included in the eighth part. The ninth part discusses how the application can be used to increase the TA percentage of a MT model. The conclusion and future works concludes the paper along with the references.

2 Related Work

NLP Tools in the field of Aviation can be traced back to TUAM AVIATION [12, 10]. Various other tools exist for pre-processing [14] and post-processing the corpus [13]. E-Dictionaries exist that help in the creation of parallel-corpus. A string of NLP tools are used individually by Boeing, Airbus and other aviation companies [10][2, 5].Though no significant work was observed for Indian languages in the field of aviation, Corpus construction for Indian languages started in the early 90s and currently is a major area of research. Some major examples of monolingual corpora for Indian languages are as follows:

Table 1: Indian Monolingual Corpus

Language used	words	Institution
Kannada, tamil	3 million	CSIL, Mysore
Assamese, Oriya	1 million	HALS, BHU
Eng, Hindi, Punjabi	3 million	IIT Delhi
Sanskrit	2 million	SS V.vidlay
Bengali	2 million	ISI kolkata

All the above mentioned projects were done under the supervision of TDIL (GOI) for major Indian languages [6]. Table 2 lists some major English-Indian language corpora in TDIL:

Table 2: Indian Bilingual Corpora

Natural Language Pair	Domain	Corpus Size
English-Assamese	Tourism	5760
English-Bodo	Agriculture	4000
English-Hindi	Health	14984
English-Urdu	Health	14860
English-Manipuri	Entertainment	10000
English-Manipuri	Agriculture	12000
English-Bengali	Tourism	11900
English-Bengali	Health	15000
English-Bengali	Agriculture	4200

Apart from the above mentioned parallel corpora, the English to Manipuri Parallel corpus was developed at CDAC, Mumbai, India [8] and English-Punjabi Parallel corpora was built at Department of CS, Punjab University, Patiala, Punjab [11]. Other worth mentioning corpora project was EMILLE/CIIL corpus that was developed by Department of Linguistics, Lancaster University, England and CIIL, Mysore, India [3] and Hindi-English and Hindi only Corpus for Machine Translation by Charles University, Prague [4].

3 The need

3.1 For the Aviation corpus:

As an attempt to create a Translation system capable of handling English-Bengali translation of aviation sentences and phrases was taken up, it was observed that there was no monolingual or bilingual corpus available for English-Indian language (pair). The unavailability of at-least a data repository for aviation sentences created a huge hurdle in the creation of the system. It was at this point that it was decided to create a bilingual English-Bengali parallel corpus where sentences in both the source and target languages would consist of aviation phrases and phraseologies.

3.2 For the Aviation Parallel corpus analyzer:

While developing the Aviation corpus, it had to be kept in mind to develop a balanced one. For it to be such it was necessary to have a detail understanding of the aviation technical terms in both the source and target language. The parallel corpus, being a uniquely technical one consisted of around 1100 out-of-vocabulary words [15]. This corpus was to be first of its kind in English-Bengali pair and in any Indian language. So care had to be taken for the corpus to be accurate. This corpus was to be the first reference in aviation domain for both native and non-native users of Bengali. The decision to develop the analyzer tool with a proper user interface was taken so as to analyze the composition of the parallel-corpus and rectify it and increase its accuracy.

3.3 For the unique vocabulary:

Chaos is avoided and uniformity maintained in the aviation world through use of a set of standard phrases implemented throughout the globe. Air Traffic Controllers, airmen and support staff uses these phrases to communicate. This aviation vocabulary in English is documented under various manuals created and implemented by ICAO and IATA. These unique phrases are an alpha-numeric composition. As a result the parallel corpus is composed of this unique vocabulary in both the source and target languages. This unique set of words arranged according to predefined English and Aviation rules form the array of instructions based on which each and every flight is accomplished. ICAO i.e. International Civil Aviation Organization devised these set of unique phrases which needs to be followed to avoid accidents. To have a better idea about these phrases we can have a look at some examples as issued by Airport authority of India through MATS (Manual of Air Traffic Services).

Table 3: AAI MATS PHRASES

Phrases	English Meaning
AFIL	An alpha character group used to designate an airfield flight plan
CPL	Current flight plan
Glide Path	Profile determined for vertical guidance during a final approach
OCA or OCH	Obstacle clearance altitude or obstacle clearance height

For satisfactory communication between all parties involved in aviation AAI has definite interpretation of words. Table 4 illustrates some of them, Table 5 consists of various ICAO addressee indicator and their meanings as used in the aviation vocabulary.

Table 4: MATS Interpretation

MATS vocabulary	Interpretation
AFIL	An alpha character group used to designate an airfield flight plan
CPL	Current flight plan
Glide Path	Profile determined for vertical guidance during a final approach
OCA or OCH	Obstacle clearance altitude or obstacle clearance height

Table 5: MATS ADRESSEE INDICATOR

Addressee Indicators	Meaning
YXY	The Case where the addressee is a military service
ZZZ	In the case where the addressee is an aircraft in flight
YYY	In all other cases

The lack of both monolingual and bilingual corpora for Indian languages in aviation makes this domain a scarily explored one. The incident and accident reports are documented by American ASRS [13], European ECCAIRS [14], and French DGAC [9] but the lack of corpora or database makes it hard to be researched, especially for Indian languages. Attempts to transliterate or translate ICAO / AAI phraseologies through standard MT systems to any Indian native languages results in failure. Standard MT systems are unable to take up the challenge to decode these phraseologies and convert them to their target language [2].

4 Building the Corpus

In order to bring about their translations in the target language the need for a proper bilingual corpus was felt. Taking English as the source and Bengali as the target language a parallel corpus was

Table 6: Features of the English-Bengali Bilingual Corpus

Domain	Natural Languages	MATS/ICAO Phraseologies used	MATS/ICAO OOV words used	Parallel Sentences
Aviation	English Bengali	700	1800	25,000

designed. The corpus was used successfully to create the first known translation system in aviation domain for any Indian language [15].The Bilingual English-Bengali corpus consists of hundreds of aviation words, phrases and phraseologies thus making it a unique one. The properties and features of the bilingual corpus as discussed in the paper [15] are as follows

5 Composition of the Bilingual Corpus

As the corpus was being built for the first time so the work constituted not only taking scarcely available sentences as references but building both the source and target sentences from the scratch. The corpus has been constituted taking the following sources in consideration [1, 7]

6 Analysing the Bilingual Corpus and Results and Discussions

As the corpus for the English-Bangla pair in Aviation domain was being created for the first time there was no other corpus in the aviation domain to compare it with. The need was felt to analyze the aviation Bilingual corpus to get a better idea of the corpus quality and also assist in Human assisted Machine Translation. The following text analysis was done on the bilingual corpus:

- Concordance
- Word list
- POS tagging
- N-Grams

In-order to design the project we used the following software: Python 3.7, NLTK and Tkinter 3.0.

7 The Proposed Methodology and Architecture

7.1 The methodology:

The working Methodology of the tool can be described through the following steps: *Step 1*: Use of the Bilingual aviation corpus (Figure 2) *Step 2*: Freezing the Structure and composition (Choice of tools to be included) *Step 3*: Deciding the architecture of the tool (Figure 3) *Step 4*: Creation of the tool and embedding in a single GUI

7.2 Approach Undertaken:

The tool was created, and embedded under the umbrella of a simple GUI, with the help of Python 3.7, NLTK and Tkinter 3.0. The tool is capable of determining the Concordance, word frequency, POS tagging and N grams of the Bilingual aviation corpus. AVRO keyboard (Figure 1) has been used for creation of Bengali sentences and UTF-8 Unicode is employed to support it.

7.3 Creating the tool:

- Python 3.7: Python coding has been used to transform the respective algorithms.
- NLTK: NLTK is array of libraries and programs has been used to support the Python programming language.
- TKinter 3.0: Tkinter is the standard Python interface to the Tk GUI toolkit. The Tkinter features that have been used for the creation of this tool are Button, Entry, Frame, Label, Menu, etc.
- UTF-8 Unicode: Unicode Transformation Format has been used for Bengali alphabets.
- AVRO keyboard: The AVRO keyboard has been used to type the Bengali sentences in the English-Bengali aviation corpus. The phonetic layout of the AVRO keyboard is as given in Figure 1.

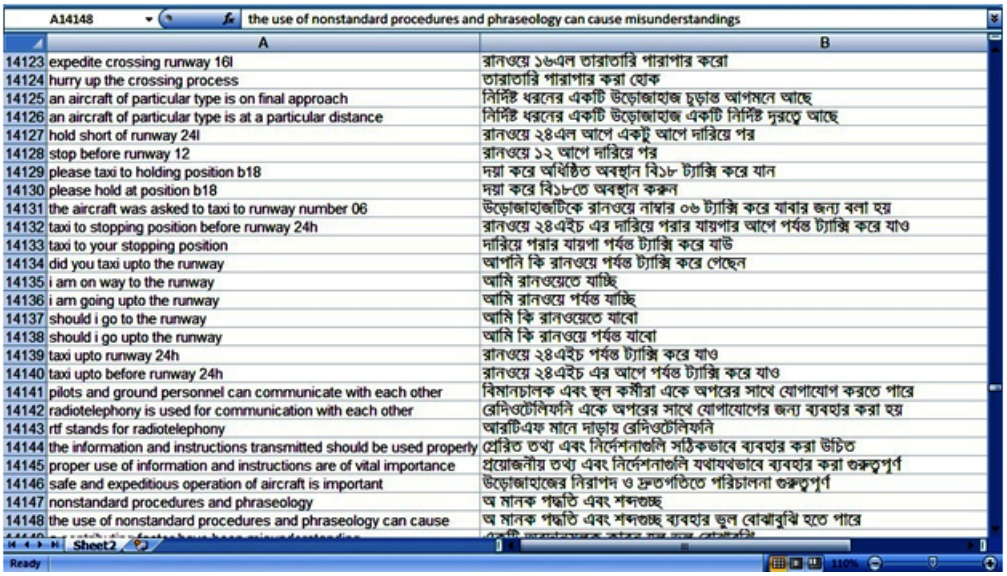
7.4 Structure and Composition of the Aviation Bilingual Corpus:

Figure. 2, shows that the bilingual corpus consists of two columns: *Column 1*: English Sentences (Source language) *Column 2*: Bengali Sentences (Target sentence)

Table 7: Sources of the Bilingual corpus

Source	**Phrases /Phraseologies**
AAI MATS 1. Document Identification and Control 2. Definitions 3. General 4. Air Traffic Service 5. Separation Methods and Minima 6. Aerodrome Vicinity procedures 7. Aerodrome control Services 8. Radar Services and Procedures 9. Flight information Services 10. Coordination 11. ATS messages 12. Phraseologies 13. ADS services 14. Controller-Pilot data link communications 15. procedures for communication failures 16. Miscellaneous Procedures 17. ATS Safety Management	890 (approximately)
Dictionary of Aeronautical English	230
FAA regulation handbook	120
ASRS Incident reports	280
DGCA reports	200
Pilot's handbook of Aeronautical knowledge	260

Figure 1: AVRO keyboard phonetic layout



	A	B
14123	expedite crossing runway 16l	রানওয়ে ১৬এল তারাতাড়ি পারাপার করো
14124	hurry up the crossing process	তারাতাড়ি পারাপার করা হোক
14125	an aircraft of particular type is on final approach	নির্দিষ্ট ধরনের একটি উড়োজাহাজ চূড়ান্ত আগমনে আছে
14126	an aircraft of particular type is at a particular distance	নির্দিষ্ট ধরনের একটি উড়োজাহাজ একটি নির্দিষ্ট দূরত্বে আছে
14127	hold short of runway 24l	রানওয়ে ২৪এল আগে একটু আগে দারিয়ে পর
14128	stop before runway 12	রানওয়ে ১২ আগে দারিয়ে পর
14129	please taxi to holding position b18	দয়া করে অর্থিষ্ঠিত অবস্থান বি১৮ ট্যাক্সি করে যান
14130	please hold at position b18	দয়া করে বি১৮তে অবস্থান করুন
14131	the aircraft was asked to taxi to runway number 06	উড়োজাহাজটিকে রানওয়ে নাম্বার ০৬ ট্যাক্সি করে যাবার জন্য বলা হয়
14132	taxi to stopping position before runway 24h	রানওয়ে ২৪এইচ এর দারিয়ে পরার যায়গার আগে পর্যন্ত ট্যাক্সি করে যাও
14133	taxi to your stopping position	দারিয়ে পরার যায়গা পর্যন্ত ট্যাক্সি করে যাউ
14134	did you taxi upto the runway	আপনি কি রানওয়ে পর্যন্ত ট্যাক্সি করে পেছেন
14135	i am on way to the runway	আমি রানওয়েতে যাচ্ছি
14136	i am going upto the runway	আমি রানওয়ে পর্যন্ত যাচ্ছি
14137	should i go to the runway	আমি কি রানওয়েতে যাবো
14138	should i go upto the runway	আমি কি রানওয়ে পর্যন্ত যাবো
14139	taxi upto runway 24h	রানওয়ে ২৪এইচ পর্যন্ত ট্যাক্সি করে যাও
14140	taxi upto before runway 24h	রানওয়ে ২৪এইচ এর আগে পর্যন্ত ট্যাক্সি করে যাও
14141	pilots and ground personnel can communicate with each other	বিমানচালক এবং ভুল কর্মীরা একে অপরের সাথে যোগাযোগ করতে পারে
14142	radiotelephony is used for communication with each other	রেডিওটেলিফনি একে অপরের সাথে যোগাযোগের জন্য ব্যবহার করা হয়
14143	rtf stands for radiotelephony	আরটিএফ মানে দাড়ায় রেডিওটেলিফনি
14144	the information and instructions transmitted should be used properly	প্রেরিত তথ্য এবং নির্দেশনাগুলি সঠিকভাবে ব্যবহার করা উচিত
14145	proper use of information and instructions are of vital importance	প্রয়োজনীয় তথ্য এবং নির্দেশনাগুলি যথাযথভাবে ব্যবহার করা গুরুত্বপূর্ণ
14146	safe and expeditious operation of aircraft is important	উড়োজাহাজের নিরাপদ ও দ্রুতগতিতে পরিচালনা গুরুত্বপূর্ণ
14147	nonstandard procedures and phraseology	অ মানক পদ্ধতি এবং শব্দগুচ্ছ
14148	the use of nonstandard procedures and phraseology can cause	অ মানক পদ্ধতি এবং শব্দগুচ্ছ ব্যবহার ভুল বোঝাবুঝি হতে পারে

Figure 2: English to Bengali Bilingual Aviation Corpus

7.5 Architecture of the tool:

The architecture of the Analysis tool is centered on the English-Bengali aviation corpus. In the 1st part (figure. 3) we can see the Text analyzer input choices through which we can access the respective function. The tool then searches the aviation corpus (sentences converted to .txt file with UTF-8 Unicode support) and the result is displayed

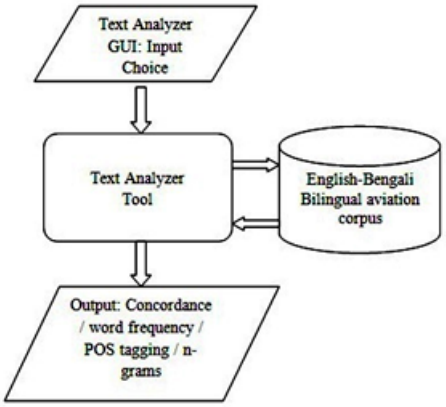


Figure 3: Overall architecture of the tool

8 Results and Output

8.1 The analyzer choice GUI:

In order to access the applicability of the Text Analysis Tool the GUI offers the choice between Concordance, Word Frequency, POS tagger and N Grams (Figure. 4). Whatever function we choose, accordingly the Bilingual corpus is used and the respective outcome is displayed.

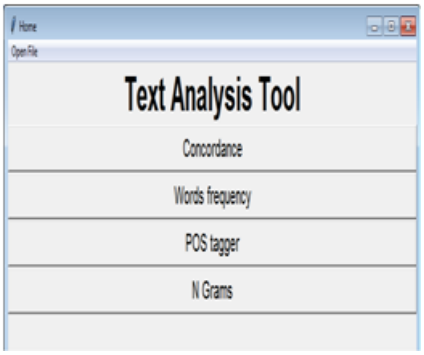


Figure 4: Text analyzer tool Interface

8.2 Concordance:

By concordance we mean the alphabetical array or list of the principal words used in corpus or data repository. Listing is done for every instance of

individual word with the respective apparent context.

Concordance tool in NLTK:

The NLTK can be used to create a simple concordance tool. The following library utilities of NLTK can be used to create the Concordance tool.

- $f = open(filename)$ It is used to open the desired file where the text that is to be analyzed is stored.
- $raw = f.read()$ This is used to read that file that is already opened.
- $tok = nltk.word_tokenize(raw)$ The function is used to tokenize all the words that are stored in the text file.
- $text = nltk.Text(tok)$ Text is used to store all the tokenized words.
- $text.concordance()$ The function concordance() produces the output with the KWIC as its argument.

For example if we choose a file named sobujsohor.txt and enter the word to be searched as "সবুজ" in Bengali, it is displayed as such:

The concordance that we are using here is Bilingual concordance. It shows all the sentences that are associated with the word in the aviation corpus and the number of times the word has occurred in the text for both the English and Bengali languages. Similarly for the word "airport" in the file airlines.txt the following output is displayed (Figure. 6). We can see that in Fig.6, total 268 matches were found for the word "airport". That is out of 25000 English sentences, 268 sentences contain the word "airport". Similarly, the use of other aviation words can also be traced in the corpus.

8.3 Word list:

In order to find the word frequency / word count the choice is available through our GUI (Figure. 7). Here, (Figure. 8, (Figure. 9, and (Figure. 10 are examples of Word list, Word Count and Cumulative word Count as deducted by our tool for the Aviation Bilingual parallel corpus

The word list, word count and cumulative word count gives us not only an idea of the aviation OOV words but also the use of all the vocabulary. The vocabulary size as determined during creation of the NMT model using the corpus was found to be 22007 unique words for English and 32265 unique words for Bangla.

In the figures (8 to 10) we see the Word list, word count and Cumulative word count of the corpus. It consists of both English and Bangla words. From the Word list, word count and Cumulative word count it has been observed how the use of the same word vary from language to language. It has been observed that while for the English aviation word "airport" has one single type application, the Bangla equivalent "বিমানবন্দর" can be used as "বিমানবন্দর" "বিমানবন্দরে" and "বিমানবন্দরটি" depending on its use. So while, the word "airport" has a count of 204, the equivalent Bangla word "বিমানবন্দর" has distribution of 57, 74 and 73 respectively. Similarly we observe that for certain English words there are two or more of Bangla equivalents depending on its use in Bangla language resulting in division of total word count among them.

8.4 POS Tagging

In order to find Parts of Speech POS tagging is used for aviation sentences, a set of Tags concerning Bangla and English has been considered. It has 24 POS tags. They are as displayed in Table X. Based on the mentioned table we enter a sentence in Bangla "আমি এখান থেকে কিভাবে এয়ারপোর্ট যাবো" and its equivalent English sentence to get the POS tagging. The results are as displayed in Figure. 11.

The equivalence of the sentence in English is "how will I go to the airport, from here". The Part of speech of the Bengali (Figure 11) and equivalent English sentence as deducted are as follows:

8.5 N-Grams:

The concept of N-gram(s) is one of the easiest to grasp in NLP. N-gram (n) means an array or sequence of "N" words. Let us take an example with respect to our aviation corpus. "Pilot is ok" is a 3 gram sentence while "the airport is near" is a 4-gram. Now in order to view the various N-gram sentences with particular aviation related word we need to enter the following data:

1. File name (airlines.txt in our case)
2. Grams to be computed (N=1, 2, 3)
3. Word to be searched (বিমান for our example)

So with respect to the word "বিমান", and N=3 the application generates the output as depicted in Figure 13. The same can be generated for any aviation

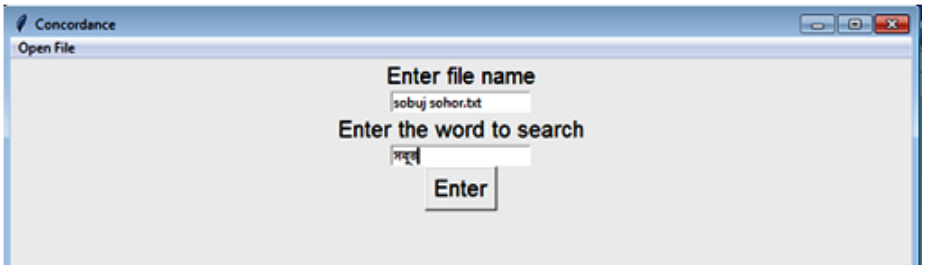


Figure 5: Finding the concordance of the word "সবুজ" in the file "sobujsohor.txt"

```
Displaying 25 of 268 matches:
মান seaplane নউবিমান airport বিমানবন্দর AI এআই 9W
ফিক কন্ট্রোল টাওয়ার AIRPORT এয়ারপোর্ট AUTOPILOT
t private GREENFIELD airport দুর্গাপুর ভারতের গ্র
 দিকে বাক নিল RAIPUR airport is the 29th busiest
 is the 29th busiest airport by passenger traffic
মানবন্দর BHUBANESWAR airport has ICAO code of VEB
h in approach to the airport বিমানটি বিমানবন্দরের
ির্দেশনা করে SILCHAR airport has DVOR DME MM ILS
মএসএসআর আছে BEGUMPET airport is equipped with all
te with ATC BEGUMPET airport বিমানচালক এসিটি বেগম
(ে যোগাযোগ করেছিল the airport at BEGUMPET is popul
y known as HYDERABAD airport বিগমপেটের বিমানবন্দর
lane movement at the airport at 0538 ০৫৩৮ এ এটিসি
I is a single RUNWAY airport মুম্বাই হল একটি রানওে
 রাডার the HYDERABAD airport is equipped with nav
GANDHI international airport এয়ার ট্রাফিক কন্ট্র
(েকসনাল (রঞ্ SILCHAR airport is equipped with ins
GANDHI international airport is situated at SHAMS
former international airport situated at BEGUMPET
GANDHI international airport is the sixth busiest
is the sixth busiest airport in India রাজীব গান্ধ
ম বিমানবন্দর SILCHAR airport is the second busies
s the second busiest airport of ASSAM শিলচর বিমান
র as a single runway airport MUMBAI airport is th
unway airport MUMBAI airport is the busiest airpo
```

Figure 6: Output of Concordance for the word "airport" in the file aviation.txt

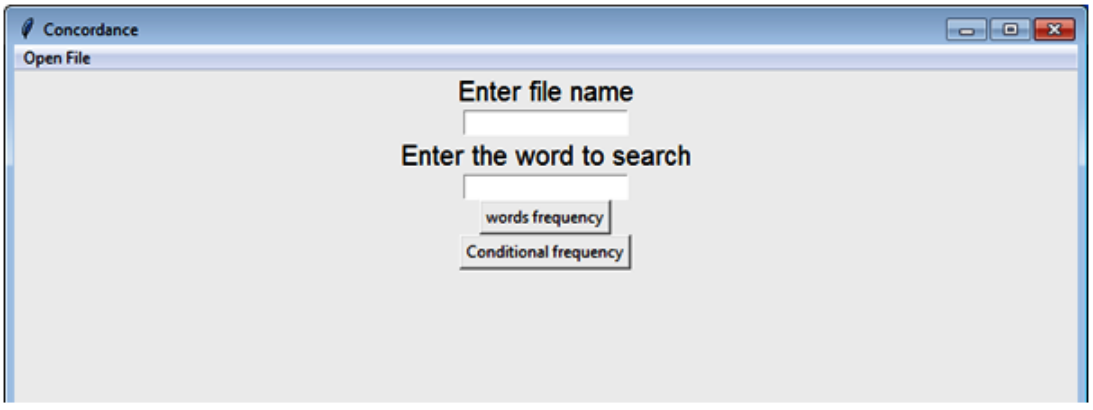


Figure 7: The interface for word count

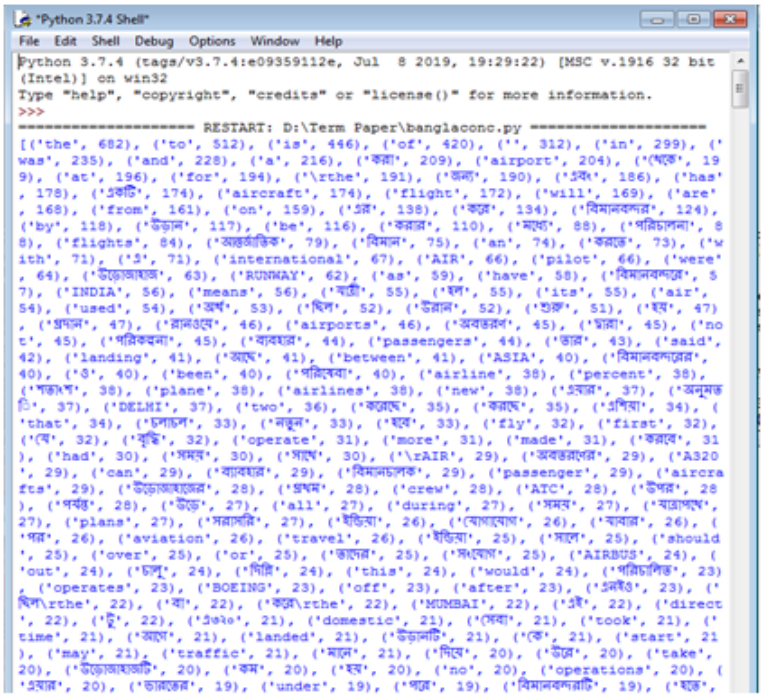


Figure 8: Word List of the bilingual corpus

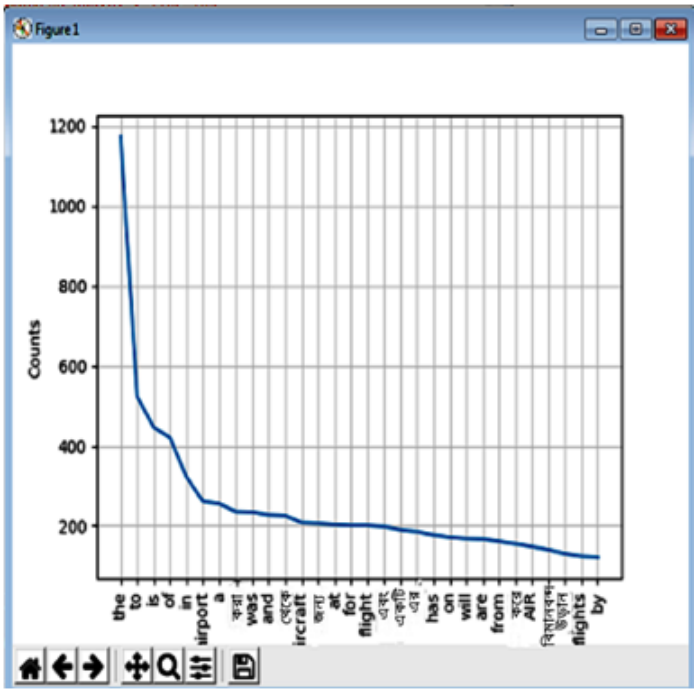


Figure 9: Word count of the Aviation corpus

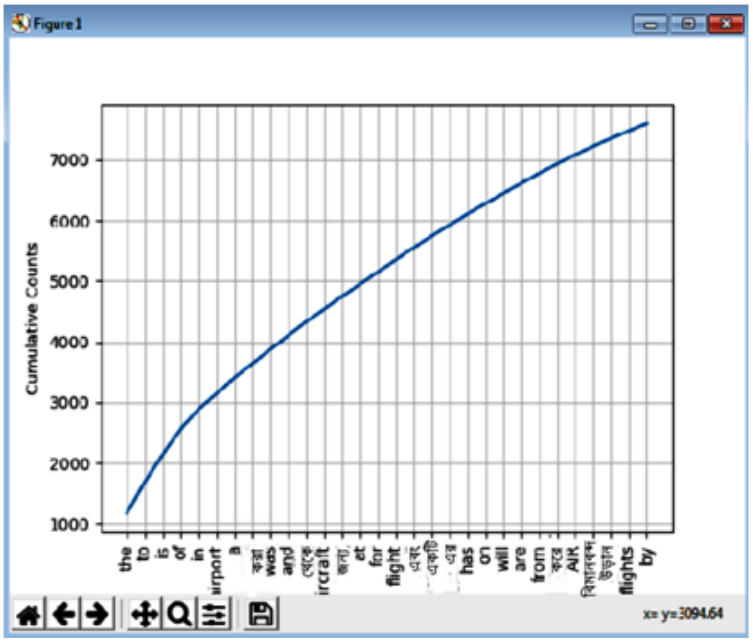


Figure 10: Cumulative word count

Table 8: Comparison of Bangla and English word usage

Word count and usage comparison of "airport" and "বিমানবন্দর"			
Word	Used as	Example	Total count
airport	Noun		204
বিমানবন্দর	Noun	This is an airport: এটা একটি বিমানবন্দর	57
বিমানবন্দরের		he is inside the airport :সে বিমানবন্দরের ভেথরে আছে	74
বিমানবন্দরটি		the airport is very beautiful :বিমানবন্দরটি খুব সুন্দর	73

Table 9: Comparison of English and equivalent Bangla words usage (Continued)

Comparison of English and equivalent Bangla words usage			
Word	Use in Bangla	Word	Used in Bangla as
Aircraft	উড়োজাহাজ	Land	অবতরণ
	উড়োজাহাজটি		
	উড়োজাহাজের		অবতরণের
	উড়োজাহাজে		

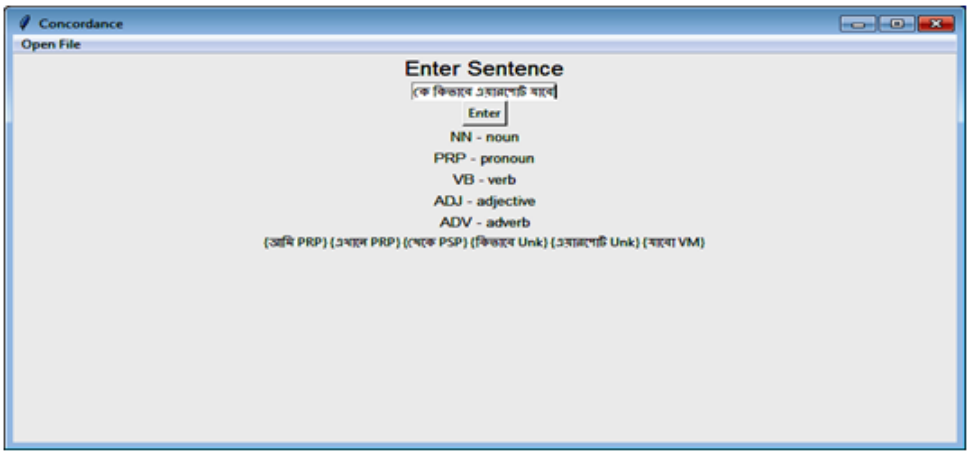


Figure 11: POS Tagging of the Bangla Sentence

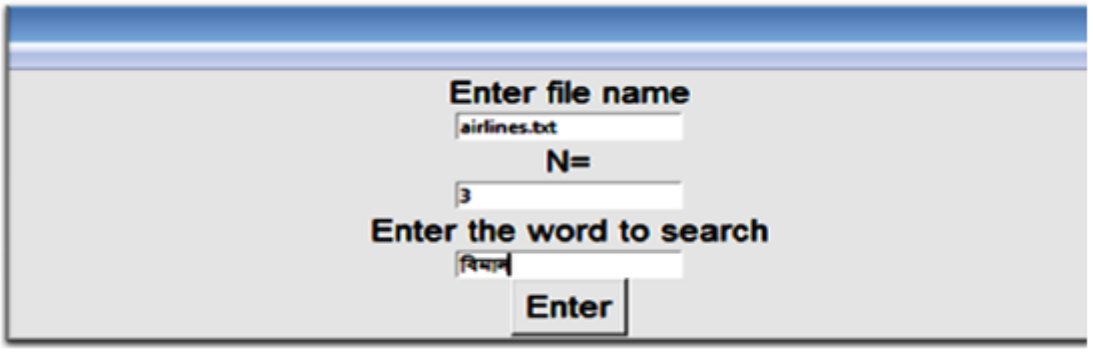


Figure 12: GUI to calculate N-Grams

Table 10: POS tags and their Description

Tag	Description
NN	Common Noun
PPP	Pronoun
NNP	Proper Noun
PSP	Postposition
JJ	Adjective
INTF	Intensifier
RP	Particles
NEG	Negative Word
RB	Adverb
QF	Quatifiers
DEM	Demonstrative
NST	Spatial Noun
SYM	Symbol
ECH	Echo Words
WQ	Question Words
QC	Cardinals
XC	Compounds
CC	Conjuncts
QO	Ordinals
RDP	Reduplication
INJ	Interjection
VM	Main Verb
VAUX	Verb Auxiliary
UNK	Unknown Words

word with an N-gram value of our requirement and research importance.

Figure 13 shows us the type of sentence combination and the number of times they appear in the corpus which contains the Bangla word "বিমান", and have n-gram value of 3. The pie chart depicts the results

9 Analysis

English to Bengali aviation corpus. We took the liberty to calculate the frequency distribution of the words used. Also it gave us an idea on the number of times various OOV words of aviation was used. This paved the way to better understand the corpus and modify and upgrade it. Lexical dispersion of the aviation words used was also done to understand the occurrence of aviation words in the corpus. Lexical dispersion on our aviation corpus explains on how frequently an OOV phrase or word appears across the length and breadth of our aviation corpus.

9.1 Frequency distribution:

While the word count (Figure 10) and cumulative word count (Figure 10) gives us an idea about the count of all words or vocabulary used in the corpus, frequency distribution is generally used to record the frequency of each word type in corpus or file. It is a more precise way to locate targeted words and their use. The frequency distribution of some aviation related English words as found in the bilingual corpus is as follows:

From Figure 15 it is evident that in our corpus the most used aviation word is "airport" followed by "aircraft", Flight, ATR and others. This type of analyzing helps us to understand the usage of particular OOV words that makes the corpus a unique one. It also helps us to determine our Translation Analysis (TA) percentage once the corpus is implemented as a Translation model.

9.2 Lexical Dispersion:

The Lexical dispersion is responsible for measuring the frequency of a particular word (Aviation OOV in our case) appearing in various parts of a corpus. It keeps track of the occurrences of out-of-vocabulary words. For our corpus we can determine the Lexical Dispersion of any words, if we consider three common words aircraft, airport and flight, then the plot is as depicted:

Ultimately the total vocabulary size (determined during creation of the model using Open-NMT) and number of OOV words and ATC phraseologies used (determined using our analysis tool) is shown in Table 12 [15]:

10 Comparision with Previous Works

The development of the aviation corpus containing English and their equivalent Bengali sentences is the first known attempt to create a corpus involving English and Indian Language (Bangla) .The results of this research work can act as a benchmark for English-Indian Languages as it was able to attain remarkable results when the NMT aviation model was implemented [15]. The closest standard corpus available is the English-Bengali Tourism corpus available in TDIL website under the ministry of Electronics and Information Technology, Govt. of India [18]. It consists of 11,977 parallel sentences. It is the only standard corpus that bears some resemblance with our Aviation corpus for English-Bengali pair. It was found though the Tourism corpus consists of, use of some similar words airport, aircraft, airlines, etc but, it lacks drastically in use of aviation OOV words like Apron, Taxiway and others. For foreign languages TUAM-AVIATION is the only know work that tried to make use of language pair concept. The aim of the project was to create English to French Machine translation system which could

Table 11: Comparison of POS of Bangla sentence and English Equivalent

Sentence in Bangla		**Equivalent Sentence in English**	
"আমি এখান থেকে কিভাবে এয়ারপোর্ট যাবো"		"how will I go to the airport, from here"	
Words (Bangla)	Part of Speech	Words (English)	Part of Speech
আমি	*Pronoun (PRP)*	how	*adverb*
এখান	*Pronoun (PRP)*	will	*verb*
থেকে	*Postposition (PSP)*	I	*pronoun*
কিভাবে	*Unknown words (UNK)*	go	*verb*
এয়ারপোর্ট	*Unknown words (UNK)*	to	*preposition*
যাবো	*Main Verb(VM)*	the	*article*
		airport	*noun*
		from	*preposition*
		here	*adverb*

{{বিমান চালনা করার} 5} {{বিমান থেকে নামতে} 3} {{বিমান পরিচালনার জন্য} 2} {{বিমান যোগ করার} 2} {{বিমান চলাচল মন্ত্রণালয়ের} 2} {{বিমান বাহিনী iaf} 1} {{বিমান seaplane লউবিমান} 1} {{বিমান অন্তর্ভুক্ত করার} 1} {{বিমান কর্মীরা ৪জন} 1} {{বিমান কর্মীরা ২} 1} {{বিমান চালনার সহায়ক} 1} {{বিমান রক্ষণাবেক্ষণের ইঞ্জিনিয়ার} 1} {{বিমান পরিচালক 9w} 1} {{বিমান উরাছিলেন pilot} 1} {{বিমান পরিবহন চালোনার} 1} {{বিমান পরিষেবা শুরু} 1} {{বিমান বাহিনী পরিবহনের} 1} {{বিমান multi role} 1} {{বিমান light transport} 1} {{বিমান হিসেবে ব্যবহৃত} 1} {{বিমান হিসেবে পরিচালনা} 1} {{বিমান antanov an32} 1} {{বিমান বাহিনীর ১১টি} 1} {{বিমান বাহিনী ১১টি} 1} {{বিমান সেবিকা a} 1} {{বিমান সেবিকা এআই} 1} {{বিমান সেবিকাটি বিষয়টি} 1} {{বিমান প্রবেশন করেছে} 1} {{বিমান যোগ করেছে} 1} {{বিমান কম i5} 1} {{বিমান বিক্রি করে} 1} {{বিমান অন্তর্ভুক্ত করবে} 1} {{বিমান জন্য ভিজিএফ} 1} {{বিমান সংস্থা উড়ান} 1} {{বিমান পরিষেবা ৩০} 1} {{বিমান all the} 1} {{বিমান বাহিনি প্রাথমিক} 1} {{বিমান বাহিনি iata} 1} {{বিমান চলাচল বাজার} 1} {{বিমান সেবিকারা বিমানে} 1} {{বিমান সেবিকারা জরুরি} 1} {{বিমান খালি করতে} 1} {{বিমান সেবিকা জরুরি} 1} {{বিমান থেকে বার} 1} {{বিমান ছিল su30mki} 1} {{বিমান চালনা বন্ধ} 1} {{বিমান runway রানওয়ে} 1} {{বিমান যলজাট the} 1} {{বিমান পরিচালনা করে} 1} {{বিমান চলাচলে বিঘ্নব} 1} {{বিমান টিকিট সংরক্ষণ} 1} {{বিমান চলাচল বৃদ্ধি} 1} {{বিমান একে অপরের} 1} {{বিমান চলাচল নিয়ন্ত্রক} 1} {{বিমান সর্বনিম্ন দূরত্ব} 1} {{বিমান অবিশ্বাস্যভাবে কাছাকাছি} 1} {{বিমান যাত্রীদের ক্রমবর্ধমান} 1} {{বিমান চলাচলে বৃদ্ধি} 1} {{বিমান বাহিনীর বিমানবন্দর} 1} {{বিমান দ্বারা যাচ্ছেন} 1} {{বিমান ভ্রমণকারীদের জন্য} 1} {{বিমান যোগাযোগ উন্নত} 1} {{বিমান যাত্রী চলাচল} 1} {{বিমান যাত্রীদের বিমানে} 1} {{বিমান আবাসিত করতে} 1} {{বিমান সংঘর্ষের পরিহার} 1} {{বিমান যলজাট হতে} 1} {{বিমান বাহিনীর বেস} 1} {{বিমান থেকে নামানো} 1} {{বিমান উপাদান এবং} 1} {{বিমান চলাচল বিঘ্নস্ত} 1} {{বিমান দ্বারা ব্যবহৃত} 1} {{বিমান ব্যবহার করে} 1} {{বিমান নিজেদের মধ্যে} 1} {{বিমান চলাচল নীতিমালা} 1} {{বিমান চলাচল জন্য} 1} {{বিমান ভাড়া উপর} 1} {{বিমান চলাচল মন্ত্রণালয়} 1} {{বিমান সি১৭ এবং} 1} {{বিমান সারা দেশ} 1} {{বিমান ভ্রমণের সময়} 1} {{বিমান ভাড়া ৩১শতাংশ} 1} {{বিমান নিদর্শনপত্র ৯০০} 1} {{বিমান পরিবহন মন্ত্রণালয়} 1} {{বিমান থাকে the} 1}

Figure 13: Output for the word "বিমান" and N=3

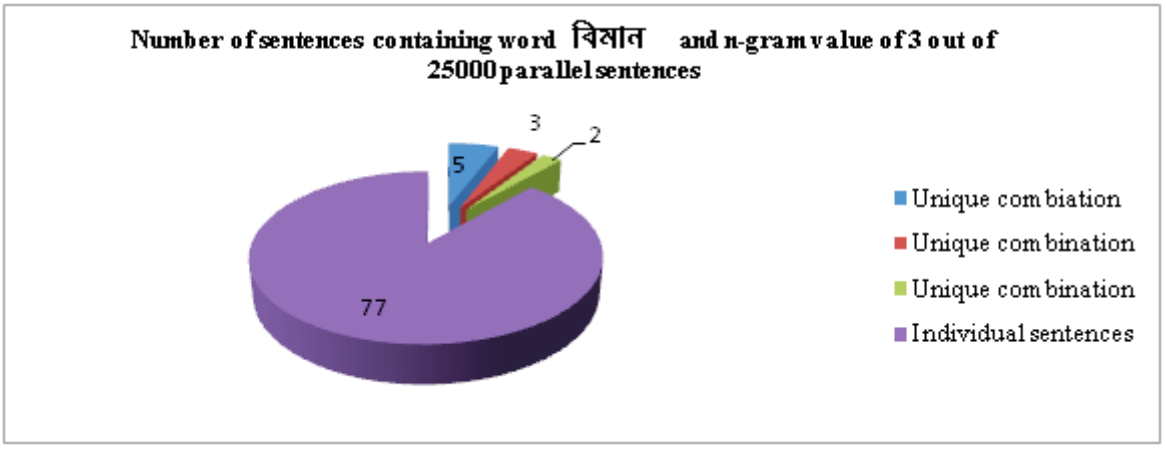


Figure 14: Aviation sentences divided by use of word "বিমান" and n-gram value of 3

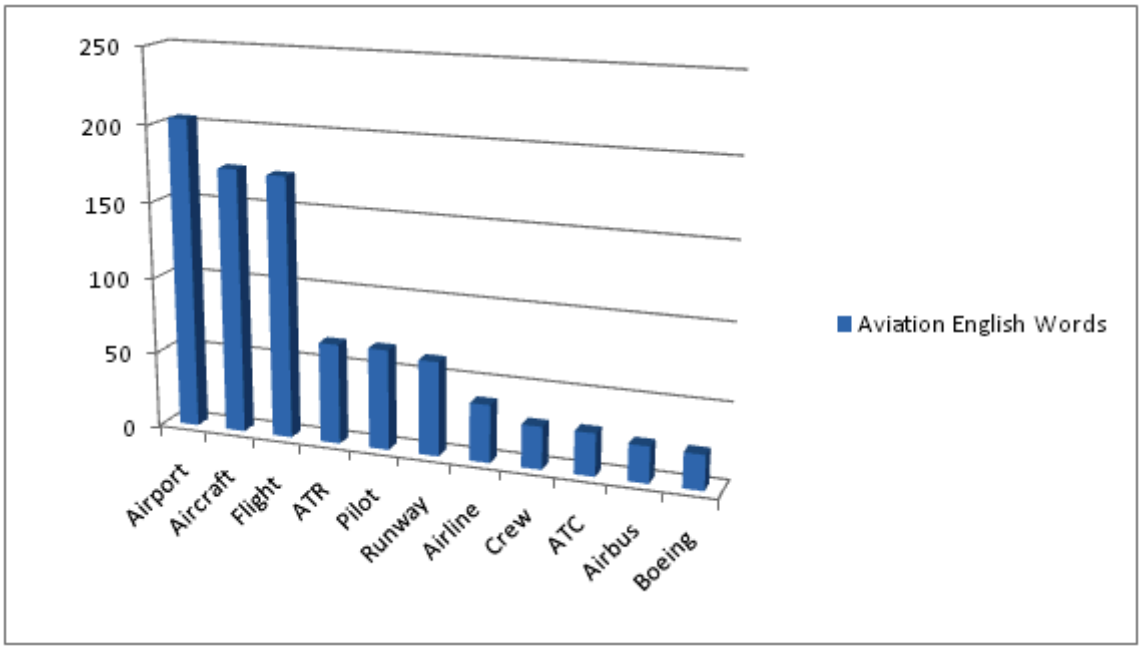


Figure 15: Frequency distribution of English aviation words

Table 12: Features of the developed Corpus

Domain	Natural Languages	Vocabulary size (OpenNMT)	ATC Phraseologies used	Aviation OOV words used
Aviation	English	22007	600	1200
	Bengali	32265		

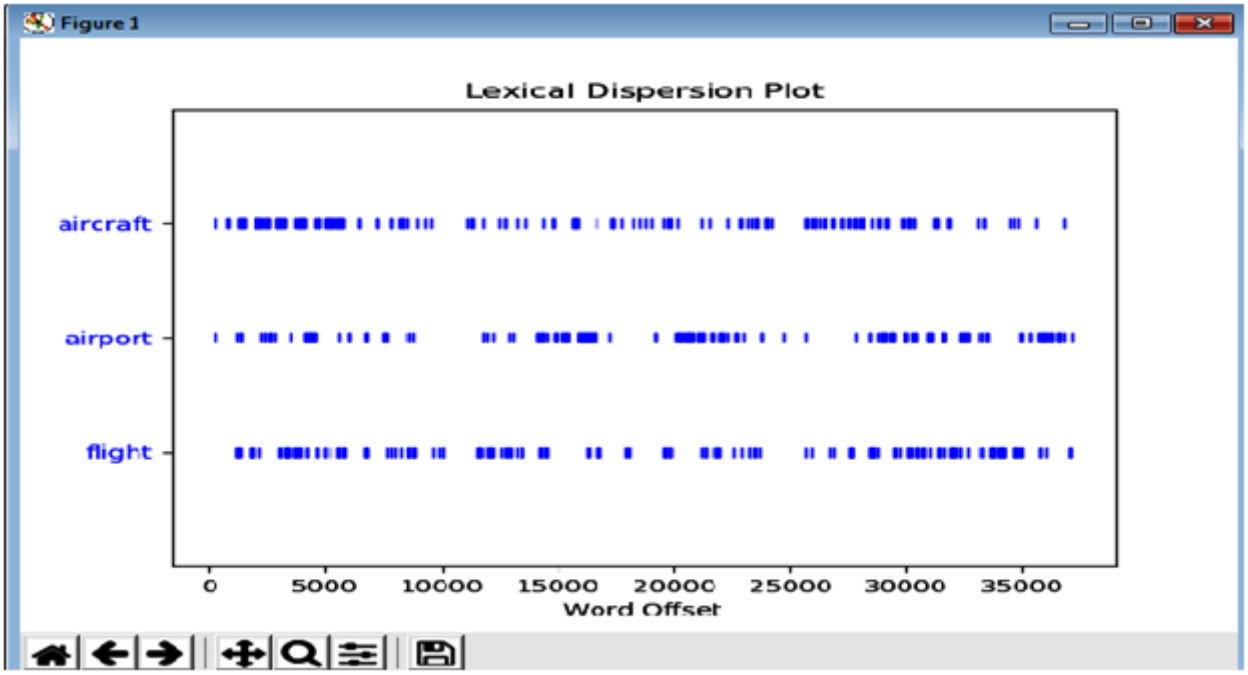


Figure 16: Lexical Dispersion Plot

translate maintenance manuals of aircrafts. In 1979 a Prototype was developed which performed translation of maintenance and hydraulic manuals. The TAUM AVIATION project was done in the following environment; Hardware: CYBER 173 and Operating System: NOS/BE 1.4 Dedicated dictionaries and grammars were used which in turn were compiled into an object code. This code was then compared and transformed at run time considering the input text. The source maintenance manuals were in English. The dictionary consisted of core vocabulary and had the following data: 4054 unique entities and the English-French Bilingual dictionary consisted of 3280 unique entities [10].

11 Use in Machine Translation:

The use of the analysis application is found when human assisted machine translation is employed. This method helps to handle the OOV terms that are not translated by the MT model. As a result the un-translated and un-transliterated words are then compared with the output of the application and appropriate steps are taken to increase the percentage of TA or translation analysis of the Model. Human assisted Machine Translation is often employed in the MT domain along with pre-processing and post processing tools, achieving satisfactory BLEU scores [15]

12 Conclusion and Future Works

The development of the Bilingual corpus has enabled us to create not only the first working English-Bengali Aviation Translation system [15] but also has paved the way for the creation of E-dictionary and other NLP tools. Though text analyzer tools have been developed for different types of text, it is the first time that it has been developed for the aviation domain. Also, the tool has been applied on English-Bengali language pair making it Bilingual text analyzer. This tool has been instrumental in increasing the TA percentage of the Neural Translation Model [15] showcasing its importance and applicability.

The array of native Indian languages presents a huge scope for researchers to work on the development of corpus and appropriate analysis tools, especially for technical and unexplored domains. The size of the corpus can be expanded in the future and its scope of coverage increased. The tool can be tried on the Tourism corpus for a comparison study. Online and android versions can be developed for better portability of the tool.

Table 13: Comparison of the developed corpus with TDIL Tourism and TUAM-Aviation

Corpus Name and Language Pair	**Source**	**Number of parallel sentences**	**Use of common Aviation words**	**Use of Dedicated Aviation Words**
Aviation (English-Bangla)	Self developed	25,000	Used frequently	Used frequently
Tourism (English-Bangla)	TDIL (Government of India)	11,977	Minimum use	none
TAUM-Aviation (English-French)	CETADOL (research centre in computational linguistics at University de Montreal)	5054 unique entities	Used frequently	Used frequently

References

[1] AAI. AAI Resources, 1996.

[2] Aerospace, A. Airbus Aerospace, 2017.

[3] Baker, P., Hardie, A., McEnery, T., and Jayaram, S. B. Constructing Corpora of South Asian Languages. In Proceeding of the EACL workshop on South Asian languages, pages 1--8, 2003.

[4] Bojar, O., Diatka, V., Rychl, P., Strank, P., Suchomel, V., Tamchyna, A., and Zeman, D. HindEnCorp - Hindi-English and Hindi-only corpus for machine translation. In Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, pages 3550--3555, 2014.

[5] Clark, P. and Harrison, P. Boeing's NLP system and the challenges of semantic representation. In Semantics in Text Processing, STEP 2008 - Conference Proceedings, pages 263--276, 2008.

[6] Dash, N. S. Language corpora: present Indian need. In Proceedings of the SCALLA 2004 Working Conference, pages 5--7. Citeseer, 2004.

[7] DGCA. DGCA reports, 1999.

[8] DorenSingh, T. Building Parallel Corpora for SMT System: A Case Study of English-Manipuri. International Journal of Computer Applications, 52(14):47--51, 2012.

[9] ECCAIRS. ECCAIRS incident reports, 2005.

[10] Isabelle, P. and Bourbeau, L. TAUM-AVIATION: ITS TECHNICAL FEATURES AND SOME EXPERIMENTAL RESULTS. Computational Linguistics, 11(1):18--27, 1985.

[11] Jindal, S., Goyal, V., and Singh, J. Building English-Punjabi Parallel corpus for Machine Translation. International Journal of Computer Applications, 180(8):26--29, 2017.

[12] Paul, S. NLP TOOLS USED IN CIVIL AVIATION: A SURVEY. International Journal of Advanced Research in Computer Science, 9(2):109--114, 2018.

[13] Paul, S. and Purkhyasta, B. S. English to Bengali Transliteration tool for OOV words common in Indian civil aviation. Journal of Advanced Database Management & Systems, 6(1):23--32p, 2019.

[14] Paul, S. and Purkhyastha, B. S. Handling aviation oov words for machine translation and corpus creation. Indian Journal of Computer Science and Engineering, 11(5):471--477, 2020.

[15] Paul, S. and Shyam Purkhyastha, B. English to Bengali Neural Machine Translation System for the Aviation Domain. Technical Report 2, 2020.

[16] Rodriguez, D. Z. and Junior, L. C. B. Determining a non-intrusive voice quality model using machine learning and signal analysis in time. INFOCOMP Journal of Computer Science, 18(2), 2019.

[17] Silva, D. H., Rosa, R. L., and Rodriguez, D. Z. Sentimental analysis of soccer games messages

from social networks using user's profiles. INFOCOMP Journal of Computer Science, 19(1), 2020.

[18] TDIL. TDIL Resources, 1999.