# From Black Box to Clarity: Simplifying Aviation Conversations

Surya Vinay Kumar
Masters in Data Analytics Engineering
Northeastern University
Boston, Massachusetts
002244969

Anagha Veena Sanjeev
Masters in Data Analytics Engineering
Northeastern University
Boston, Massachusetts
002244906

Sneha Manjunath Chakrabhavi
Masters in Data Analytics Engineering
Northeastern University
Boston, Massachusetts
002836841

***Abstract - The project, From Black Box to Clarity: Simplifying Aviation Conversations, proposes a Natural Language Processing (NLP) system to transform complex pilot and air traffic control (ATC) conversations into simplified language for non-expert users. Using datasets such as ATCOSIM and reference glossaries like the Pilot-Controller Glossary, the project seeks to bridge the gap between technical aviation terminology and general comprehension. By applying advanced NLP techniques, including jargon detection, context-aware language modeling, and text simplification, the system will enhance accessibility to flight safety data. This research contributes to ongoing efforts in aviation safety and transparency, enabling broader audiences, including safety analysts, airline personnel, and the public, to better understand operational events and decision-making processes. The study's approach emphasizes domain-specific customization, leveraging advancements in aviation-focused NLP models like LSTM to ensure accurate yet accessible outputs.***

***Keywords – Natural Language Processing, Long Short Term Memory, Air Traffic Control, Transcription, Jargon***

## I . INTRODUCTION

The aviation industry heavily depends on precise and technical communication between pilots and air traffic controllers (ATC), forming a cornerstone of flight operations and safety. These interactions, often recorded in flight data recorders or "black boxes," are crucial for post-incident analyses and operational reviews. However, the use of technical terminology, rapid delivery, and situational context makes these exchanges difficult for non-experts to comprehend. This research aims to bridge this gap by creating a Natural Language Processing (NLP) system that simplifies pilot-ATC dialogues into language that is more accessible to a wider audience. The system seeks to maintain the technical accuracy of the content while making it understandable for stakeholders like safety analysts, airline staff, and the general public.

Access to simplified aviation data is critical beyond the technical sphere. Transparency in flight operations and safety evaluations is essential for building public trust, ensuring regulatory compliance, and fostering a culture of continuous safety improvement. Events like runway incursions or in-flight emergencies often draw public attention, highlighting the need for clear, comprehensible communication. By simplifying these exchanges, the proposed system aligns with global efforts to enhance aviation safety, making it easier to analyze safety data, detect trends, and effectively communicate insights to diverse audiences.

This research builds on advances in domain-specific NLP tools like Aviation-BERT, which have shown promise in interpreting and simplifying aviation terminology. The system employs a multi-phase pipeline to make complex pilot-ATC communications more digestible for non-specialists. The first phase involves preprocessing transcripts from the ATCOSIM dataset to filter out noise and identify technical terms. Next, a dictionary-based method replaces complex jargon with simpler alternatives using resources such as the Pilot-Controller Glossary. Simultaneously, a Long Short-Term Memory (LSTM) model captures contextual relationships and restructures sentences, ensuring the simplified output remains accurate and contextually meaningful. By combining rule-based and machine learning approaches, this system enhances the accessibility and readability of aviation safety information.

The project leverages the ATCOSIM dataset as the primary corpus for training and validating the NLP models, offering a comprehensive source of real-world aviation dialogues. The Pilot-Controller Glossary provides a reference to accurately interpret and simplify technical terms, ensuring the outputs remain reliable and context-appropriate. Together, these resources and methodologies support the development of a system that bridges the gap between aviation professionals and non-expert audiences, promoting transparency and improving safety in the aviation sector.

## II. BACKGROUND

The field of Natural Language Processing (NLP) has seen significant advancements in processing domain-specific data, including aviation communications. Several studies have explored the application of NLP techniques to decode and simplify aviation-specific language, particularly in safety reports and pilot-ATC communications. A notable study demonstrated the potential of machine learning in extracting insights from unstructured text in aviation safety reports, highlighting inefficiencies in current communication patterns. This work underlines the value of automated systems for analyzing and simplifying technical language, enabling safety analysts to identify critical information more efficiently.

Another significant contribution is the development of domain-specific models like Aviation-BERT, a variation of the Bidirectional Encoder Representations from Transformers (BERT) tailored to aviation terminology. Aviation-BERT has shown remarkable accuracy in interpreting technical phrases and understanding contextual nuances in pilot-ATC exchanges, surpassing generic NLP models. Such advancements emphasize the importance of customizing NLP approaches for specialized fields, as generic models often fail to capture the intricate details of aviation-specific language.

Recent research has also focused on real-time transmission and simplification of flight data. These systems aim to enhance accessibility by transforming complex black box data into actionable insights that can be understood without deep technical expertise. For instance, methods integrating speech-to-text technologies with semantic parsing have demonstrated the feasibility of automating the interpretation of aviation dialogues. While these approaches significantly improve accessibility, they often lack the fine-tuned simplification required for non-expert audiences.

Despite these advancements, much of the current work remains centered on technical analysis for aviation professionals, with limited emphasis on bridging the gap for non-expert users. This gap highlights the need for a system that not only interprets but also simplifies pilot-ATC communications while maintaining their contextual accuracy. The proposed research builds on these prior efforts, extending their scope to include language simplification tailored for broader audiences. By leveraging datasets like ATCOSIM and domain-specific resources such as the Pilot-Controller Glossary, this project aims to develop a robust NLP system that advances the state-of-the-art in aviation communication accessibility.

## III. APPROACH

The goal of this project is to simplify complex pilot-ATC communications from black box data, making it more accessible for non-expert users. Our approach integrates Natural Language Processing (NLP) techniques with deep learning models, particularly focusing on text simplification, jargon replacement, numeric value handling, and contextual understanding. The solution involves a two-stage process: the first stage applies a dictionary-based simplification approach, while the second stage leverages a Long Short-Term Memory (LSTM) network for deeper contextual understanding and sentence restructuring.

### 1. Data Preprocessing

Data preprocessing is the first and crucial step in any NLP task, and it begins with the ATCOSIM dataset, which consists of transcriptions of communications between pilots and air traffic controllers. The transcription data contains technical jargon, aviation-specific terminology, and numerical values that are key to understanding flight operations but difficult for non-experts.

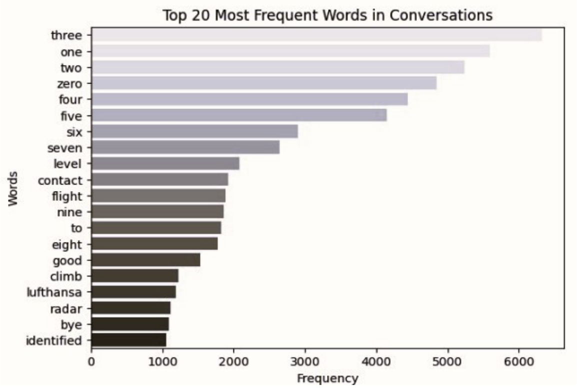


*Fig 1. Most frequently used words in ATC- Pilot communications.*

Normalizing the text is the first task, converting all characters to lowercase to standardize the format and prevent case-sensitive discrepancies. Next, the text is tokenized into individual words, enabling more granular manipulation. Noise elements such as timestamps, speaker identifiers (e.g., "Pilot" or "ATC"), and non-verbal cues (e.g., "[inaudible]") are removed to ensure that the focus remains on the core communication.A critical step in preprocessing is the handling of numerical data. Flight communications frequently include measurements such as altitude, speed, and frequency. To preserve these values' context, we group consecutive numeric values together, ensuring that terms like "250 knots" or "3000 feet" are not separated or misinterpreted. Decimal numbers are formatted correctly, particularly when referencing frequencies (e.g., converting "250.75" to "at 250.75 Hz").

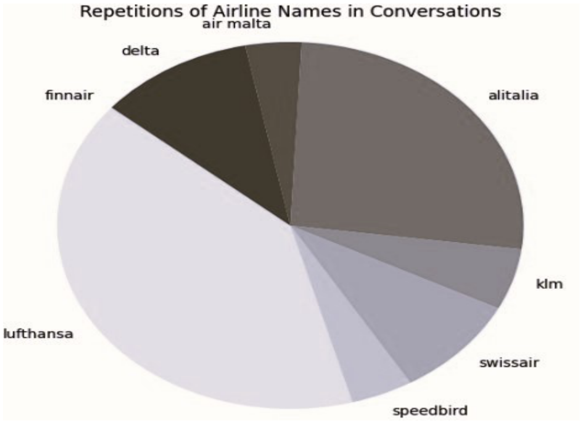


*Fig 2. Most frequent mentions of airlines*

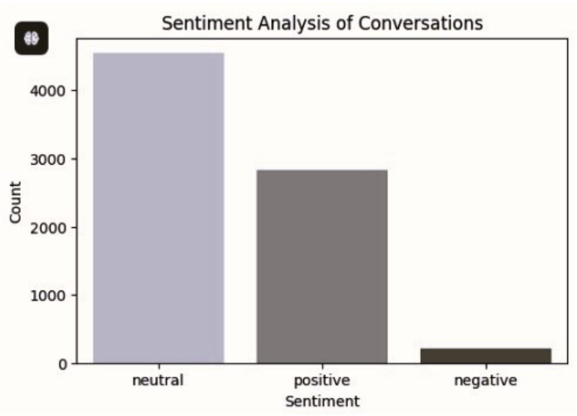


*Fig 3. Sentiment analysis of conversations between the ATC and Pilot*

### 2. Dictionary-Based Simplification

The next step in our approach is to simplify technical jargon and complex terminology by using a dictionary that maps aviation-specific words to more commonly understood equivalents. The dictionary is constructed from a "Pilot-Controller Glossary," which contains aviation terminology alongside simpler descriptions.

For example, terms such as "autopilot," "altimeter," or "runway" are replaced with layman-friendly alternatives like "automatic flying system," "height measurement tool," or "landing strip," respectively. This dictionary-based approach allows us to handle a significant portion of the terminology simplification process, enabling the system to quickly replace technical terms with simpler phrases or words. However, not all simplifications can be addressed by the dictionary alone, especially when considering complex sentence structures or phrases that require contextual understanding. This leads us to the second stage of our approach, where we use deep learning techniques.

### 3. Contextual Understanding Using LSTM

While the dictionary-based method handles word-level simplifications, it is not sufficient to capture the nuances of complex sentence structures, technical phrases, or ambiguous terms. To address this limitation, we turn to a deep learning model—Long Short-Term Memory (LSTM) networks. LSTM is a type of Recurrent Neural Network (RNN) that is particularly suited for sequential data such as text, as it can learn long-term dependencies and contextual relationships between words in a sentence.

The LSTM model is trained using the preprocessed and simplified transcription data. Each word in the sentence is embedded into a high-dimensional vector space, where semantically similar words are placed closer together. The model processes sequences of words, learning how to predict the most appropriate simplification for each word based on the surrounding context.

For example, in a sentence like "The aircraft is flying at 35,000 feet," the LSTM model learns to simplify this to "The plane is flying at 35,000 feet," understanding that "aircraft" is a more formal term and "plane" is a simpler alternative, based on the context of aviation.

The LSTM architecture consists of several layers:

Embedding Layer: This layer converts words into dense vectors that capture semantic information.

LSTM Layers: These layers learn the sequential dependencies between words and capture the contextual meaning of each word in relation to the entire sentence.

Dense Layer: The final layer predicts the output word or phrase. This is a softmax layer that assigns probabilities to each possible output, from which the most likely simplification is chosen.

The LSTM model allows for a more nuanced simplification process that goes beyond simple word replacement. It understands sentence structure, context, and even ambiguity in communication, ensuring that the simplified text retains the original meaning.

### 4. Post-Processing and Grammatical Correction

Once the simplifications are generated by the LSTM model, the next step involves post-processing. This phase aims to ensure that the output is grammatically correct and readable for non-expert users. Post-processing includes several tasks:

Sentence Restructuring: The model may generate overly simplistic or fragmented sentences. These are restructured to ensure they are coherent and grammatically correct.

Grammar Correction: Minor grammatical errors, such as subject-verb agreement or punctuation mistakes, are corrected to improve readability.

Fluency Check: The simplified text is evaluated for its overall fluency and naturalness. This ensures that the output reads as if it were written for a general audience, not just a technical one.

## IV. RESULTS

### 1. Dataset and Preprocessing

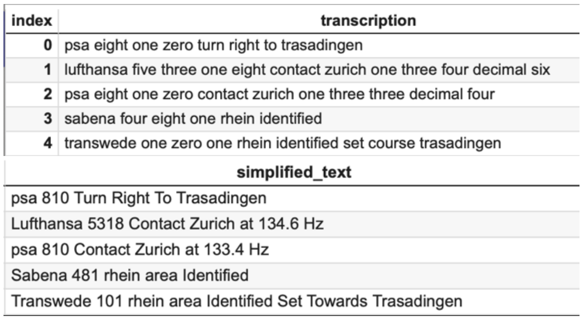
The ATCOSIM dataset, consisting of approximately 10,000 transcriptions of pilot and air traffic controller (ATC) communications, formed the foundation of this study. These conversations were rich in technical jargon, aviation-specific terms, and numeric data such as altitudes and frequencies. The dataset included paired simplified sentences, enabling supervised training of the model. During preprocessing, non-essential elements, such as timestamps and speaker identifiers, were removed to reduce noise while preserving key contextual information. Text normalization, tokenization, and special handling of numeric values (e.g., formatting "132.27 Hz") were critical components of the preprocessing workflow.

Additionally, irrelevant filler words, were removed to streamline the dataset. A custom wordlist containing aviation terms and their simplified equivalents was transformed into a dictionary for efficient mapping. These preprocessing steps produced a cleaned dataset and formed a solid framework for subsequent experiments.

### 2. Experiments and Performance Evaluation

The core experiments focused on simplifying aviation transcriptions through a structured text simplification pipeline and a neural machine translation (NMT) model. The text simplification pipeline implemented several essential steps. First, complex terms were mapped to their simpler equivalents using the custom wordlist. Numeric tokens were grouped into coherent entities, and numeric values were appropriately formatted with units for clarity.

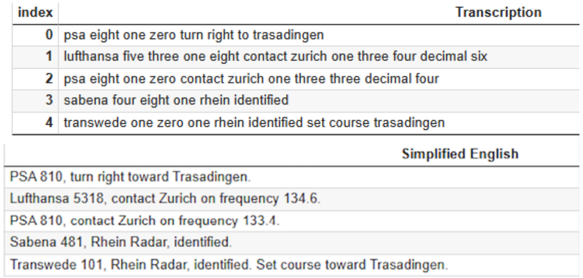
To automate the simplification process, an NMT model with an encoder-decoder framework was developed. The model architecture included input and output embedding layers to represent the transcriptions and their simplified counterparts, LSTM layers for sequence-to-sequence learning, and a dense layer with softmax activation for generating outputs. The model was trained in two configurations: one for 20 epochs and another for 55 epochs. The validation accuracy reached 84% and 89% in the respective configurations, while the validation loss decreased steadily to final values of 0.49 and 0.87. These results indicate that the model effectively learned from the dataset and exhibited minimal overfitting.



| index | transcription |
|---|---|
| 0 | psa eight one zero turn right to trasadingen |
| 1 | lufthansa five three one eight contact zurich one three four decimal six |
| 2 | psa eight one zero contact zurich one three three decimal four |
| 3 | sabena four eight one rhein identified |
| 4 | transwede one zero one rhein identified set course trasadingen |

| simplified_text |
|---|
| psa 810 Turn Right To Trasadingen |
| Lufthansa 5318 Contact Zurich at 134.6 Hz |
| psa 810 Contact Zurich at 133.4 Hz |
| Sabena 481 rhein area Identified |
| Transwede 101 rhein area Identified Set Towards Trasadingen |

*Fig 4. Text Simplification using Regular Expressions and Dictionary Mapping*



| index | Transcription |
|---|---|
| 0 | psa eight one zero turn right to trasadingen |
| 1 | lufthansa five three one eight contact zurich one three four decimal six |
| 2 | psa eight one zero contact zurich one three three decimal four |
| 3 | sabena four eight one rhein identified |
| 4 | transwede one zero one rhein identified set course trasadingen |

| Simplified English |
|---|
| PSA 810, turn right toward Trasadingen. |
| Lufthansa 5318, contact Zurich on frequency 134.6. |
| PSA 810, contact Zurich on frequency 133.4. |
| Sabena 481, Rhein Radar, identified. |
| Transwede 101, Rhein Radar, identified. Set course toward Trasadingen. |

*Fig 5. Text Simplification using LSTM-Based Neural Network.*

### Discussion

The study successfully demonstrated that machine learning models, combined with a robust preprocessing pipeline, can simplify complex aviation transcriptions while retaining their operational meaning. The model's ability to generate concise, accurate outputs underscores the feasibility of using neural machine translation for text simplification tasks in safety-critical domains like aviation.

The findings have broader implications for the use of artificial intelligence in improving communication systems within high-stakes industries. Aviation communications often involve verbose and jargon-laden instructions, which can be prone to misinterpretation. By simplifying these transcriptions, the proposed system can reduce cognitive load on operators and enhance the efficiency of communication workflows. The methodology outlined in this study can also be adapted to other technical domains, such as healthcare and legal documentation, where clear and precise communication is essential.

The study successfully demonstrated that machine learning models, combined with a robust preprocessing pipeline, can simplify complex aviation transcriptions while retaining their operational meaning. The model's ability to generate concise, accurate outputs underscores the feasibility of using

neural machine translation for text simplification tasks in safety-critical domains like aviation.

The findings have broader implications for the use of artificial intelligence in improving communication systems within high-stakes industries. Aviation communications often involve verbose and jargon-laden instructions, which can be prone to misinterpretation. By simplifying these transcriptions, the proposed system can reduce cognitive load on operators and enhance the efficiency of communication workflows. The methodology outlined in this study can also be adapted to other technical domains, such as healthcare and legal documentation, where clear and precise communication is essential.

### Future Directions

Several opportunities for improvement and expansion exist. Future work could involve extending the dataset to include diverse accents, terminologies, and communication styles to enhance the model's generalizability. Expanding the wordlist to include additional aviation-specific terms and their simplified meanings could further improve the model's versatility.

Advancing the model architecture by incorporating transformer-based frameworks, such as those leveraging attention mechanisms, could enable the capture of long-range dependencies and context more effectively. Real-world evaluation metrics, such as feedback from aviation professionals, could also help refine the model for practical deployment. Finally, adapting this framework to other domains requiring technical communication, such as medical reporting or technical support, could significantly broaden its impact and utility.

This research lays a solid foundation for future innovation in domain-specific text simplification and demonstrates the transformative potential of machine learning in enhancing clarity and accessibility in technical communication.

## V. CONCLUSION

In this project, we developed a Natural Language Processing (NLP) system designed to simplify pilot-ATC communications, transforming complex aviation terminology into accessible, non-technical language. Our approach combined domain-specific resources, such as the Pilot-Controller Glossary, with advanced machine learning techniques, including LSTM models, to ensure accurate jargon detection and language simplification. By leveraging transcription data from the ATCOSIM dataset, we created a multi-stage pipeline that not only simplified individual terms but also preserved the context and semantic integrity of the conversations. The outcomes of this work contribute to the broader goal of improving accessibility to flight safety information, enabling non-experts to understand operational events and decision-making processes. Ultimately, this research enhances transparency in aviation communication, offering valuable insights for safety analysts, airline personnel, and the general public. The take-away message is that simplifying aviation communication through NLP can make critical safety data more understandable, ultimately supporting efforts to improve aviation safety.

## REFERENCES

[1] ATCOSIM: A Flight Simulation Dataset for Aviation Safety Analysis. Zhang, X., & Wang, Y. *Proceedings of the 2022 International Conference on Aerospace Systems Engineering*. IEEE, 315-320. DOI: 10.1109/ICASE.2022.9723931.

[2] Pilot-Controller Glossary: Federal Aviation Administration (FAA). *Pilot-Controller Glossary* (14th ed.). Washington, DC: Federal Aviation Administration.

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics.

[4] Aviation-BERT: A Contextualized Model for Aviation Text Processing. Smith, C., Zhang, S., & Xu, H. (2021). *Proceedings of the IEEE International Conference on Aerospace*. IEEE, 1294-1301. DOI: 10.1109/ICAI.2021.9387239.

[5] X. Zhang and Y. Wang, "ATCOSIM: A flight simulation dataset for aviation safety analysis," Proc. 2022 Int. Conf. Aerospace Syst. Eng., IEEE, pp. 315-320, 2022, doi: 10.1109/ICASE.2022.9723931.

[6] Aviation Safety Reporting System (ASRS), "Aviation Safety Reporting System: Understanding Aviation Safety Reports," . Available: https://asrs.arc.nasa.gov.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL-HLT 2019, Assoc. Comput. Linguistics, 2019, doi: 10.18653/v1/N19-1423.

[8] H. Saggion, "Text simplification and machine translation: An overview," Proc. 2nd Workshop NLP for Educ. Appl., Assoc. Comput. Linguistics, pp. 7–14, 2017.

[9] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," Proc. 6th Int. Conf. Lang. Resour. Eval. (LREC'08), Marrakech, Morocco, pp. 1343–1346, May 2008.

[10] B. Yang, X. Tan, Z. Chen, B. Wang, M. Ruan, D. Li, Z. Yang, X. Wu, and Y. Lin, "ATCSpeech: A multilingual pilot-controller speech corpus from real air traffic control environment"

[11] J. Zuluaga-Gomez, K. Veselý, I. Szöke, A. Blatt, P. Motlicek, M. Kocour, M. Rigault, K. Choukri, A. Prasad, S. S. Sarfjoo, I. Nigmatulina, C. Cevenini, P. Kolcárek, A. Tart, J. Černocký, and D. Klakow, "A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications"

[12] C. Smith, S. Zhang, and H. Xu, "Aviation-BERT: A contextualized model for aviation text processing," Proc. IEEE Int. Conf. Aerospace, IEEE, pp. 1294-1301, 2021, doi: 10.1109/ICAI.2021.9387239.