Aviation English Corpus: A prospectus

Motivation for the study: Why is an Aviation English Corpus important?

With rapidly expanding air traffic worldwide, the demands on air traffic control communication (ATCC) are constantly increasing. ATCC often involves interaction between pilots and air traffic controllers with different linguistic backgrounds and training. In the busy airspace, efficient and accurate communication between pilots and controllers is crucial for ensuring safe operation. Recognising the importance of clear communication, several measures are in place in ATCC. Above all, the communication between controllers and pilots is highly standardised in order to minimise the potential for misunderstanding and to make the communication more efficient. The standardisation exists at the level of communication strategies (e.g. the use of readback) as well as at the level of vocabulary and grammar (so called standardised phraseology). The central role of language has been recognised also by the International Civil Aviation Organisation (ICAO), which since 2008 requires pilots and air traffic controllers to demonstrate sufficient proficiency both in standard phraseology and 'plain language' for safe communications in non-routine and emergency situations (ICAO 2010). Despite the importance of communication for aviation safety, there is a lack of research that would systematically examine the use of language by pilots and air traffic controllers. We propose to address this gap by building a large corpus (collection of transcripts and texts) of aviation English that would enable a systematic study of language use in air traffic communication and of the factors playing a role in successful information exchange. Findings based on the corpus will enable us to improve the communication and provide a basis for language training and assessment of pilots and air traffic controllers.

A brief overview of previous research: What has been done so far?

So far, the research on aviation communication has drawn on four main sources: a) accident/incident reports and cockpit voice recordings (e.g. Sassen 2005); b) pilot-controller communication elicited from flight simulators (e.g. Molesworth & Estival 2015; Burki-Cohen & Kendra 2001; Prinzo 1998); c) experimental (laboratory) studies (e.g. Barshi & Farris 2003) and d) naturally-occurring (live transmission) data (e.g. Prinzo 2009; 2006). The first three sources have provided valuable insights into different aspects of air traffic communication and helped to highlight potential difficulties. However, these data sources also have several limitations. Simulated and experimentally-elicited communication cannot fully substitute observation of the complex, authentic environment in which ATCC takes place; accident/incident reports and cockpit voice recordings provide data from real events but they focus predominately on language occurring at the time of an incident/accident (e.g. the cockpit voice recordings only provide recordings of up to two hours before the accident occurred) and do not show the broader context of the communication in which the event was embedded. Likewise, these data sources do not provide information about whether the type of communication that occurred during the incident/accident is common or not in ATCC.

To deal with these limitations, these data sources can be complemented by findings from naturally-occurring (routine as well as non-routine) air traffic communication between pilots and controllers. Several studies recognised the value of such data and transcribed recordings of transmissions between pilots and controllers (see the Appendix for a list of the largest corpora of spoken aviation English). Unfortunately, these corpora provide a very fragmentary picture of ATCC as most of the studies had a very specific (narrowly-defined) focus which guided the selection of the data for the corpora. This may significantly limit their usefulness for investigation of other issues in ATCC. Finally, the vast majority of these corpora were collected in the 1990s and thus reflect the situation before the ICAO language proficiency requirements.

Despite the limitations, the findings from these studies have demonstrated the great potential of corpus-based approach to the investigation of ATCC. For example, with only twenty-four hours of recordings from Toronto and Dublin airports, Hinrich (2008) was able to demonstrate a systematic variation in the use of standard phraseology used by controllers and pilots. Her study of question use by pilots and controllers highlighted important strategies for clarification and information-seeking. The study also pointed to potential limitations of standardised phraseology in this area. In an earlier study based on about fifty hours of recordings from US airports, Cardosi (1993) identified patterns in ATC that resulted in clear recommendations for training of pilots with respect to seeking of clarifications in communication with the tower.

Corpus linguistics: What is our approach to aviation English?

We want to approach the study of aviation English from the *corpus linguistics* perspective. Corpus linguistics works with large databases of language (language corpora) using sophisticated quantitative methods to identify and visualise patterns of language use. In this project, we can draw on existing infrastructure and support from the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University, one of the world's leading centres for corpus research, which has recently been awarded the Queen's Prize for Higher Education. The centre has extensive experience with building large spoken corpora of native and non-native English. For example, currently two major spoken corpora are being developed at the centre: BNC2014 Spoken, the spoken component of the new British National Corpus and the Trinity Lancaster Corpus, the largest spoken corpus of English produced by non-native speakers.

Data: Which interactions will be collected?

We will be seeking to obtain large amounts of data from different stages of air traffic (air-ground) communication, containing routine as well as non-routine communication. The recordings from the period following the implementation of the ICAO language proficiency requirements will be collected for the corpus. Additional data from other types of communication (both spoken and written) may also be collected. Metadata (i.e. background information) about the speakers will also be collected.

Areas of research: What will be investigated?

Our primary aim is to create a large (annotated) corpus that allows systematic (automated) research of important patterns of communication (including communication errors) in aviation with the focus on factors such as the pilot's/controller's background, language background and proficiency, personal style of communication, experience etc. One of the advantages of the corpus approach is the fact that it allows addressing a large number of research questions. The research based on the corpus can inform the following areas:

- Accident prevention
- Pilot, controller and other staff training
- Aviation English testing and test development
- Teaching materials development
- Linguistic and communication studies in general

In this project we are thus seeking to create a unique resource that will contain systematically-collected, robust evidence about interaction in aviation, allowing the linking of different types of interactions with information about speakers, context of interaction and other relevant variables.

References

Barshi, I., Farris, C. (2013). *Misunderstandings in ATC communication: Language, cognition, and experimental methodology*. Burlington, VT: Ashgate.

Burki-Cohen, J., & Kendra, A. (2001). Air traffic control in airline pilot simulator training and evaluation. *Air Traffic Control Quarterly*, 9(3), 229-253.

Cardosi, K. M. (1993). *An analysis of en route controller-pilot voice communication.* Cambridge, MA: John A. Volpe National Transportation Systems Center.

Godfrey, J. (1994). *Air Traffic Control Complete* LDC94S14A. Philadelphia: Linguistic Data Consortium.

Hinrich, S. (2008). *The use of questions in international pilot and air traffic controller communication*. (Phd Thesis). Oklahoma State University, U.S.

International Civil Aviation Organization. (2010*). Manual on the Implementation of ICAO Language Proficiency Requirements (DOC 9835 AN/453). Second edition*. Montreal: International Civil Aviation Organization.

Molesworth, B., & Estival, D. (2015). Miscommunication in general aviation: The influence of external factors on communication errors. *Safety Science*, *73*, 73-79.

Morrow, D., Lee, A., & Rodvold, M. (1993). Analysis of Problems in Routine Controller-Pilot Communications. *International Journal of Aviation Psychology*, 3 (4), 285–302.

Pavlinović, M., Boras, D., & Francetić, I. (2013). First steps in designing air traffic control communication language technology system: Compiling spoken corpus of radiotelephony communication. *International Journal of Computers and Communications, 3* (7), 73-80.

Prinzo, O.V. (1998). *An analysis of voice communication in a simulated approach control environment.* Oklahoma City, OK: FAA Civil Aeromedical Institute (NITS No. DOT/FAA/AM-97/17).

Prinzo, O.V., Hendrix, A.M. & Hendrix, R. (2006). *The outcome of ATC message complexity on pilot readback performance*. Federal Aviation Administration Report DOT/FAA/AM-06/25.

Prinzo, O.V., Hendrix, A.M. & Hendrix, R. (2009). *The outcome of ATC message length and complexity on en route pilot readback performance*. Federal Aviation Administration Report DOT/FAA/AM-09/2.

Sassen, C. (2005). *Linguistic dimensions of crisis talk: Formalising structures in a controlled language*. Amsterdam/Philadelphia: John Benjamins Publishing.

Šmídl, L. (2012). *Air traffic control communication corpus*. Published in LINDAT/CLARING repository, available under CC BY-NC-ND 3.0 from http://hdl.handle.net/11858/00-097C-0000-0001-CCA1-0

APPENDIX

Corpus name and reference	**Recordings**	**Period covered**	**Location of recordings**	**Purpose of collection**	**Availability**
Air Traffic Control Communication Speech Corpus (Šmídl, 2012)	140 hrs (planned)	not stated	Czech Rep., Lithuania, the Philippines	automatic speech recognition research	20 hours available
The Air Traffic Communication corpus (Godfrey, 1994)	70 hrs (about half contains speech)	1994, 1997	US (three locations)	automatic speech recognition research	available for a charge
Prinzo et al. (2006)	51 hrs	2003, 2004	US (five locations)	research on factors affecting language use in ATCC	not known
Prinzo et al. (2009)	50 hrs	2006	US (five locations)	research on factors affecting language use in ATCC	not known
Cardosi (1993)	47 hrs	1990s	US (different locations)	research on air communication, with focus on errors	not known
Morrow et al. (1993)	42 hrs	1990s	US (different locations)	research on factors affecting language use in ATCC	not known
Pavlinović, Boras & Francetić (2013)	40 hrs (selective episodes)	2012, 2013	Croatian airports	research on designing ATCC language technology system	not known
Hinrich (2008)	24 hrs	2005(assumed)	Toronto, Dublin	research on the use of questions in aviation English	not known