First Steps in Designing Air Traffic Control Communication Language Technology System - Compiling Spoken Corpus of Radiotelephony Communication

Mira Pavlinović, Damir Boras, and Ivana Francetić

***Abstract*—**One of the most essential parts of air traffic control is communication. It helps air traffic controllers and pilots operate the plane and maintain safe and expeditious flight. A survey of the NASA Aviation Safety Reporting System database has identified that lack of radiotelephony communication skills and discipline by pilots and controllers is a causal or circumstantial factor in 80% of incidents or accidents. The goal of this paper is to provide an overview of spoken corpus of radiotelephony phraseology recorded on the frequencies of Zagreb Approach and Tower Control. The spoken corpus of radiotelephony communication has been compiled and will be used as a basis for designing a language technology system that should spot deviations from the prescribed usage of radiotelephony communication. The recordings have been made during peak hours of traffic at Zagreb Airport Pleso. Out of recorded forty hours, twenty hours (ten hours of communication from Zagreb Approach Control and ten hours from communication on Tower Control) have been selected to be transcribed and incorporated into the language technology system. Although the designed corpus of spoken radiotelephony communications was recorded in Croatian airspace and is relatively small, it is found to be representative for radiotelephony language used in Europe and therefore applicable to any research in European radiotelephony communications.

***Keywords*—**air traffic control communication, approach and tower control, radiotelephony communication, spoken corpus of radiotelephony communication, language technology system

I. Introduction

To maintain the highest level of safety in aviation, a lot of research has been done to improve air traffic control and pilot operational systems. Voice communication system used by air traffic controllers and pilots is the only segment that has not been developed and improved in the course of the past fifty years. In today's crowded airspace with a global steady growth in air traffic, it is important that communication is performed in a standardised and understandable way to all air traffic participants. Constant insisting on proper usage of radiotelephony phraseology results in automated usage of communication procedures and contributes to air traffic safety. Any deviation from the standardized phraseologies presents an obstacle to the best possible communication. To maintain the highest level of safety, International Civil Aviation Agency (ICAO) prescribed strict rules that govern communication between a pilot and controller. The rules for this language, radiotelephony phraseology, are located in Annex 10, Volume II, and Chapter 12 of Doc 4444 and further explained and implemented by national service providers. In Croatia this is done by Croatia Control Ltd. It sets the communication system architecture that provides fast, safe and reliable flow of information between aircraft in the controlled airspace and Air Traffic Control (ATC) centres as well as between Croatian and foreign ATC centres. As ICAO standardized phraseology is not fully harmonized on a worldwide basis every states publish differences with respect to ICAO Standards. Croatia Control Ltd., Aeronautical Information Service, issues Radio Communication Procedures (Voice Communication in Aeronautical Mobile Service) in a document called AIC. The Croatian radiotelephony phraseology, technique, and procedures are based on ICAO (Standards and Recommended Practices).

This paper gives an outline of the spoken radiotelephony communication corpus that was compiled and created as one segment of a doctoral study research. The compiled corpus will be used for setting up the language technology model that should spot deviations from the standard usage of prescribed radiotelephony communication. The reasons for compiling the corpus as well as its main advantages, constrains and possibilities for further research and application are further provided in the article.

M. Pavlinović is with the Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva 4, Zagreb, Croatia (corresponding author to provide e-mail: mira.pavlinovic@fpz.hr).

D. Boras is with the Department of Information and Communication Sciences, University of Zagreb, Faculty of Humanities and Social Sciences, Ivana Lučića 3, Zagreb, Croatia (e-mail: dboras@ffzg.hr).

I. Francetić is with the Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva 4, Zagreb, Croatia (e-mail: ivana.francetic@fpz.hr).

II. Communication

The role of the Air Traffic Control is to ensure safe, orderly and expeditious flow of traffic. One of the most essential and vital tasks that air traffic controllers and pilots perform is communication. Air traffic controllers give instructions, issue clearances and guide pilots through the airspace by means of voice communication.

Communication is defined as an exchange of information, ideas and knowledge. A traditional model of communication, proposed by Shannon and Weaver (Fig. 1), is a system in which an information sender and the receiver are required. The sender encodes his or her intended meaning in a spoken utterance and transmits it to the receiver. The utterance is conveyed via the appropriate channel in the form of a sound-stream which is perceived and decoded by the receiver. The receiver's representation of the meaning of the utterance will, in the case of successful communication, be a perfect or near-perfect match of the sender's intended meaning. This model of spoken verbal communication is addressed by the ICAO language proficiency requirements.

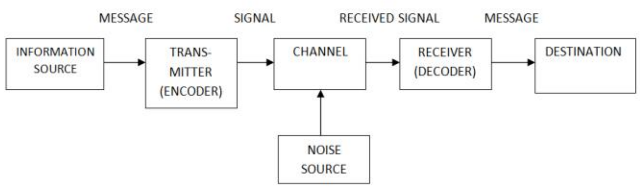


Fig. 1. Traditional model of communication.

In Air Traffic Control it is of crucial importance that all parties involved in communication understand each other, that there is no place for ambiguities and misunderstanding, and that the information is delivered and received timely and accurately.

One of the deadliest accidents in aviation history, which resulted in 583 fatalities, was a collision involving two Boeing 747 passenger aircraft at the Tenerife airport in 1977. It was a defining event in aviation safety and a tragic lesson in communication. This accident demonstrated that information transmitted by radio communication can be understood in a different way to that intended. The KLM airplane was in position and holding and the co-pilot asked for a takeoff clearance. The air traffic controller gave the clearance but did not explicitly say they were cleared for take-off. Reading back the clearance, the co-pilot stated that they were *Taking off* without using the prescribed phrase for that situation *Cleared for take-off*. When the controller replied with the words *Oka'* the pilots understood this as a clearance for taking off. While KLM was on the take-off roll, the Pan American plane and the controllers both radioed at the same time, cancelling each other's calls that the KLM should not take off yet. KLM did not hear the radio call and continued resulting in a crash that killed hundreds.

Due to many factors such as call sign confusion, readback/hearback error, noise, open microphones, number problems, ambiguity, expectation, etc., the oral transmission of essential information through a single and vulnerable radio frequency carries many potential dangers.

III. Air Traffic Control Structure

According to the European Organisation for the Safety of Air Navigation (EUROCONTROL), air traffic controllers are responsible for guiding aircraft through the airspace safely and efficiently. The goal of Air Traffic Control is to minimize the risk of aircraft collisions while maximizing the number of aircraft that can fly safely in airspace at the same time. Aircraft pilots and their on-board flight crews work closely with controllers to manage air traffic [2]. The pilots flying the aircraft through the airspace are obliged to precisely follow the instructions of the air traffic controllers. Air Traffic Control is a combination of four general elements:

a. The first element is the basic set of flying rules that pilots follow in the air.

b. The second element is the multitude of electronic navigation systems, landing system and instruments that pilots use.

c. The third element is the division of airport surface and air space in different type of control areas. Air traffic controllers operating in each of these areas and the computer systems they use to track aircraft during take-off, landing and in flight are also part of this element.

d. The fourth element is the communication between pilots-controllers, controllers-controllers and the equipment used for this communication [2].

Communication between a pilot and an air traffic controller synchronises what air traffic controller decides and utters and what pilot does with an airplane. The controller monitors the plane and gives instructions to the pilot. As the plane leaves that airspace division and enters another, the air traffic controller passes it off to the controller or controllers responsible for the new airspace division.

Every flight is divided into seven different phases (Fig. 2): pre-flight, take-off, departure, en-route, descend, approach, and landing. Each phase is defined by what the plane does and is handled by a different controller.

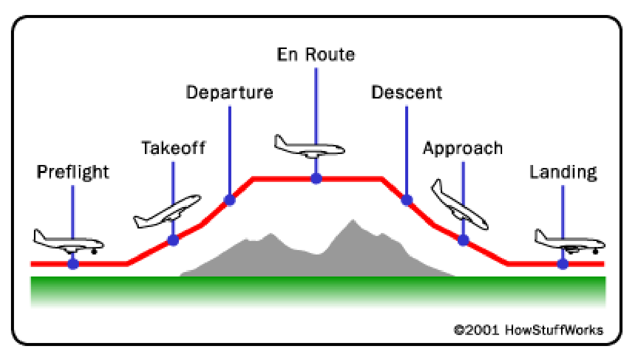


Fig. 2. Phases of flight [13].

As defined in Annex 11 to the Convention on International Civil Aviation, Objectives of Air Traffic Services are:

• to prevent collisions between aircraft in the air and on the manoeuvring areas of aerodromes

• to prevent collisions between aircraft and other vehicles and obstructions on the manoeuvring area of aerodromes

• to maintain a safe, orderly and expeditious flow of air traffic taking into consideration the abatement of avoidable noise

• to provide advice and information useful for the safe, orderly and expeditious conduct of flights

• to notify appropriate organisations regarding aircraft in need of search and rescue and to assist such organisations as required. Division of Air Traffic Services is shown in Fig. 3.

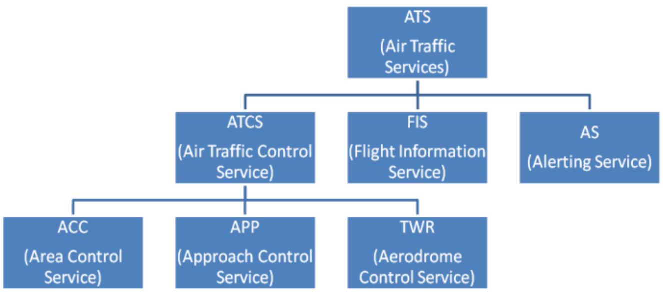


Fig. 3. Division of Air Traffic Services.

The proposed language technology system will be done for Approach and Tower Control and voice communication has been recorded on the frequencies of the mentioned controls. The Approach Control is a unit established to provide air traffic control service to controlled flights arriving at, or departing from one or more aerodromes.

Approach Control handles:

- departing aircraft
- arriving aircraft and
- overflights.

Functions of Approach Control are:

- to provide separation
- to maintain an expeditious flow of air traffic
- to assist pilots to avoid areas of adverse weather
- to assist pilots with navigational problems
- to issue traffic information
- to help pilots in special situations (emergencies, search and rescue, flight-tests, calibration flights, etc...).

The responsibility of the Tower Control is to ensure that sufficient runway separation exists between aircraft landing and departing.

IV. Radiotelephony Phraseology

Use either Radiotelephony phraseology is standardised means of communication by which pilots and air traffic controllers communicate and represents a set of operational procedures. It is an organised system for transmission of information, advice, instructions, clearances and permissions from the sender to the receiver and from the receiver to the sender. Radiotelephony phraseology is a set of prescribed rules that define what to needs to be said in a certain situation, when and how to say that, and finally how to understand the uttered.

It is a restricted and coded sublanguage with reduced vocabulary in which each word has a precise meaning that is often exclusive to the aviation domain. It is carried out in English, but the meaning of standardised phrases differs a lot from their meaning in plain English. For example: *Monitor* means *Listen out on (frequency), Out* means *This exchange of transmission is ended and no response is expected, Cleared* means *Authorised to proceed under conditions specified*, etc.

The construction of sentences also differs from plain English: sentences are short, there are no determiners (the, my, his, etc.), no auxiliary verbs (is, are), no modal verbs (may, might, can, could, etc.), no subject pronouns (I, we, you, they, etc.), and many prepositions are removed. The vast majority of sentences are either in passive or imperative. Here are some examples of radiotelephony language:

Proceed with your message.

Hold short of runway.

Taxi to holding point.

Line up and wait.

The usage of standard radiotelephony phraseology facilitates a common understanding among speakers and reduces the possible ambiguities of spoken language. In order to ensures effective communication and decrease the number of communication errors pilots and air traffic controllers are obliged to perform communication in a specific communication loop:

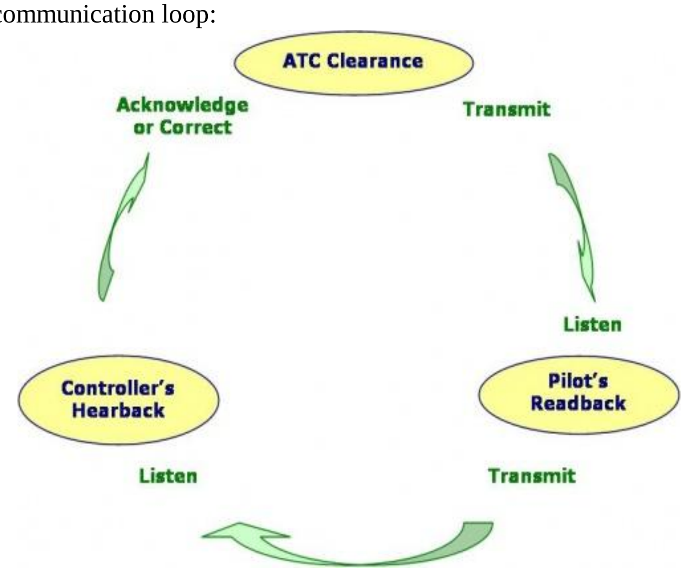


Fig. 4. Pilot – controller communication loop[15].

1. The controller utters an instruction or a clearance through a headset system.

2. The instruction is transmitted through a satellite network to the pilot.

3. The pilot than receives the instruction using the headset and replies back.

Pilots always have to read back instructions received from air traffic controllers and controllers have to listen to the readbacks and confirm them.

The mandatory items listed below have to be read back fully by the pilot:

a) Taxi/Towing Instructions
b) Level Instructions
c) Heading Instructions
d) Speed Instructions
e) Airways or Route Clearances
f) Approach Clearances
g) Runway-in-Use
h) Clearance to Enter, Land On, Take-Off On, Backtrack, Cross, or Hold Short of any Active Runway
i) Secondary Surveillance Radar Operating Instructions
j) Altimeter Settings
k) VHF Information
l) Frequency Changes
m) Type of ATS Service
n) Transition Levels [11].

The pilot will request from the controller to repeat or clarify the instruction or clearance that is not fully understood and if the controller does not receive the readback, the pilot will be asked to do that.

According to a survey carried out by the NASA Aviation Safety Reporting System (ASRS), 80 % of incidents or accidents are caused by incorrect or incomplete pilot/controller communications (Table 1). Incorrect communication, absence of communication, and correct but late communication are recognised and identified as factors affecting pilot/controller communication.

Table I. Factors affecting pilot/controller communication [11].

Factor	Percentage of Reports
Incorrect Communication	80%
Absence of Communication	33%
Correct but late Communication	12%

Incorrect or inadequate communication such as air traffic control instructions (e.g. radar vectors, heading instructions, altitude), weather or traffic information and advice/service in case of emergency are stated to be causal factors in more than 30 % of approach-and-landing accidents [11].

Readback / hearback errors may result in one or more of the following types-of-event, ranked by number of events observed over the period 1992-1993:

- Operational deviation (non-adherence to legal requirements);
- Altitude deviation;
- Airborne conflict;
- Less than desired separation;
- Lateral deviation;
- Runway incursion;
- Ground conflict;
- Airspace penetration; and,
- Near midair-collision [11].

There are several reasons why *non-standard phraseology* is a major obstacle to effective communications:

1. Standard phraseology in pilot-controller communication is intended to be universally understood.

2. Standard phraseology helps lessen the ambiguities of spoken language and thus facilitates a common understanding among speakers:

(a) Of different native languages; or,

(b) Of the same native language, but who use, pronounce or understand words differently.

3. Non-standard phraseology or the omission of key words may completely change the meaning of the intended message, resulting in potential traffic conflicts.

4. For example, any message containing a number should indicate what the number refers to (e.g. flight level, heading or airspeed). Including key words prevents erroneous interpretation and allows an effective readback/hearback.

5. Particular care is necessary when certain levels are referred to because of the high incidence of confusion between, for example, FL100 and FL110.

6. Non-standard phraseology is sometimes adopted unilaterally by national or local air traffic services, or is used by pilots or controllers in an attempt to alleviate these problems; however, standard phraseology minimises the potential for misunderstanding [15].

According to the previous researches types of miscommunication can be grouped as follows:

1. Absent-mindedness and Slips
2. Ambiguity
3. Callsign Confusion
4. Code Switching
5. Different Voices
6. Emergencies
7. Enunciation
8. Expectation
9. Headsets
10. Homonyms and Homophony
11. Noise
12. Not Hearing
13. Number Problems
14. Open microphones
15. Readback Error
16. Similarity of SIDs (Standard Instrument Departures), STARs (Standard Recommendations and Practices) and Waypoints
17. Speech Acts
18. Speed of Delivery and Pauses
19. Vigilance [15].

The proposed language technology system, in which transcripts from the compiled spoken corpus of the radiotelephony communication on Zagreb Approach and Tower frequency are included, is firstly meant to be used for training purposes and later might be further developed and used in real air traffic control communication. Most air traffic control simulation facilities use the pseudo pilot concept to simulate the communication with aircraft pilots. Each controller working position is equipped with a radio communication link to pseudo pilots in an adjacent room. The pseudo pilots listen to the clearances and enter the relevant parameters via a terminal which is connected to the simulation process.

V. Language Technology System

The proposed language technology should make communication between air traffic controller and pilot more efficient and reliable and could contribute to the increase in safety of aviation. The system could be used not only to support the pilot/controller communication, but to assist with training.

The system should be used for detecting two groups of problems:

1. language-based communication problems (unfamiliar RT phraseology, incomplete or incorrect readback/hearback utterances); c
2. communication problems with numbers (altitude, heading, etc.).

This proposed system is meant to be tested and applied for the Approach and Tower Control Unit as, according to the interviews with the instructors of RT Communications at the Faculty of Transport and Traffic Sciences in Zagreb and air traffic controllers, the largest portion of communication between a pilot and an air traffic controller takes place during these phases of flight.

The functionality of this language system will be described using scenarios to demonstrate communication within the Approach and Tower Control, and will be demonstrated using *Wizard of Oz* usability test. In *Wizard of Oz* approach, the subject acts as a "user" and interacts with the system, presumed to be a computer. It assumes that the resulting "dialogues" will be typical of human-machine interaction in their nature. This would appear to run against the aim of permitting the user to talk to the system in as natural a manner as possible, i.e. of requiring the interface to mimic human-human dialogue as well as possible [6].

Scenarios are a software definition method developed by Carroll and associates. The simplest description is that they are stories that provide a common ground for all stakeholders in a software development team to understand the functionality of the system. The focus is primarily put on the user. They give a context of a plot with actors and the events that lead towards a certain goal or objective. Thinking about the functionality this way compels the designers of the system to look at the rationale for the functionality and to focus on the use of the system. The end result is a fixed interpretation on the functionality that is being designed over the technology being used. The scenarios that describe the situation will be defined and it will be shown what will change if the language technology system is introduced in the system. For example:

ATC gives the following clearance to the pilot: "*Zagreb Control, CTN 751, with you overhead 60 north 40 west at 0803, flight level 340*."

Pilot replies: "*CTN 751, 60 north 40 west at 0803, flight level 390*."

The language technology system compares the readback with the clearance and discovers the discrepancy between what the pilot said and what the controller cleared. The language technology system warns the controller by sending the following text message to the screen: "*Warning! Flight level incorrect.* "

The findings and results *Wizard of Oz* usability test will serve as guidelines for designing a fully functional language technology system. The system will consist of:

1. Radiotelephony corpus
2. Automatic speech recognition software
3. Speech-to-text software
4. Extraction software
5. Text warning on the screen.

VI. Collection, design and analysis of spoken corpus of radiotelephony phraseology

The mentioned language technology system ad created spoken corpus of radiotelephony phraseology are a part of doctorial study research. The goal of that research is to look into communication flow within air traffic control services, and to develop and propose a language technology system that could spot deviations from the usage of standard phraseology and warn about incorrect readbacks.

A. *Corpus collection*

The first step in designing the language technology model was compilation of radiotelephony communication corpus. The first idea was to compile a corpus with all instructions and clearances listed in Radio Communication Procedures (Voice Communication in Aeronautical Mobile Service) published by Croatia Control. After listening live communication on the frequency, it was realised that around 40% of communication differs from the prescribed communication, but even as such it is widely used and understood. So, it was decided that messages frequently used and accepted as valid by pilots and controllers in live radio communication will be included in the construction of the model. If only standardized radiotelephony were used as basis for setting the model, the model would the majority of time report on incorrect utterances and would not be functional.

Therefore, it was decided to make recordings of live radiotelephony communication, extract phrases that are frequently used and recognised as acceptable, and compile a corpus that consists of recorded phrases and prescribed radiotelephony phraseology.

The recordings used for corpus design were collected during November and December 2012 and January 2013 on the frequencies of Zagreb Approach Control (120.7 MHZ) and Zagreb Tower Control (118.3 MHZ). Icom VHF air band transceiver IC-A24, Omnidirectional Base Station Antenna CXL 3-1LW and a laptop were used for making recordings. Icom IC-A24, a device that receives and transmits radio waves for the 118 - 137 MHz civil aircraft band, reduces noise caused by atmospherical discharges was connected to outdoor base station antenna. The received signal was recorded on the laptop by a Goldwave commercial digital audio editor software and stored as mp3 files. Mp3 files do not require a lot of storage memory and are easy to handle and process.

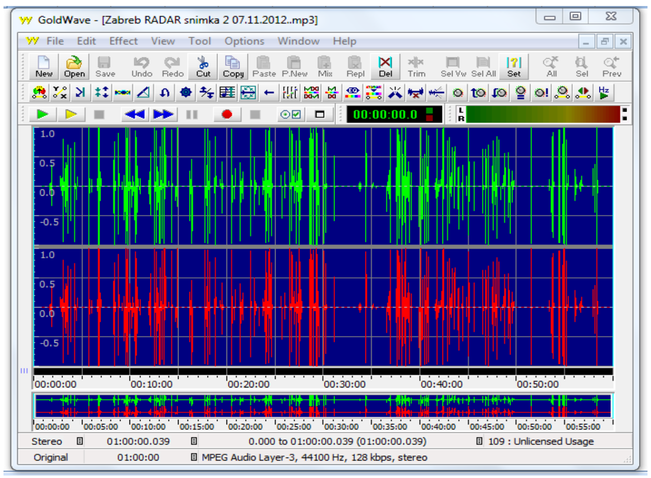


Fig. 5. Goldwave user interface for the recording made on 7th November 2012.

Although the equipment used for corpus recording is suitable for recording in noisy surrounding and difficult conditions, some recordings were quite demanding for transcription due to bad reception, noise and occasional interruptions in the receipt of signal. It was needed for some recordings or parts of recordings to be played numerous times in order to understand communication. It took approximately five to six hours to transcribe one hour of the recorded communication.

Forty hours of communication were recorded on Zagreb Approach Control (120.7 MHZ) and Zagreb Tower Control (118.3 MHZ) frequency.

The recording were made during peak hours of traffic at Zagreb Airport Pleso, i.e. during morning hours (form 8.00 to 11.30), middle of the day (from 14.30 to 17.00), and evening hours (from 19.00 to 22.30). Taking into consideration the quality of the recordings and traffic density, out of recorded forty hours, twenty hours (ten hours of communication from Zagreb Approach Control and ten hours from communication on Tower Control) were selected to be transcribed.

B. *Corpus design*

The corpus is designed from three different groups of data. The first group consists of 556 standard radiotelephony phrases prescribed by Radio Communication Procedures. The second group is designed from transcripts of the recordings and contains 1967 exchanges. The number of exchanges relates to the number of massages exchanged between a pilot (P) and a controller (C). An extract from a conversation from 7th November 2012 contains five messages:

C: *Lufthansa One Papa Hotel you will be number two reduce speed to two five two five zero knots.*

P: *Reducing two fifty Lufthansa One Papa Hotel.*

C: *Lufthansa One Papa Hotel descend to flight level one zero zero.*

P: *Lufthansa One Papa Hotel descending level one hundred.*

C: *Lufthansa One Papa Hotel from present position fly on heading two two zero maintain flight level one zero zero.*

The third set of contains terminology used at airports (e.g. names of airport vehicles, names of runways and taxiways at Zagreb airport, etc.) information relevant for Croatian airspace (waypoints, routes, etc) and information on procedures that are carried out at Zagreb airport. Although the last set of data is relevant only for Croatian airspace, the users of Croatian airspace are of various nationalities and it can be stated that the set of collected phrases is representative for radiotelephony language used in Europe.

C. *Corpus analysis*

When the recordings were collected, transcripts made, and necessary information collected, the radiotelephony corpus was designed and analysed with Oxford WordSmith Tools 04 (2007). Oxford WordSmith Tools 04 is a set of linguistic tools used for determining how words behave in a text. It consists of several tools: WordList, Keywors and Concorde tool. The system requirements needed for using these tools are an average computer with Windows 2000 or later and texts saved as plain text (.txt) file.

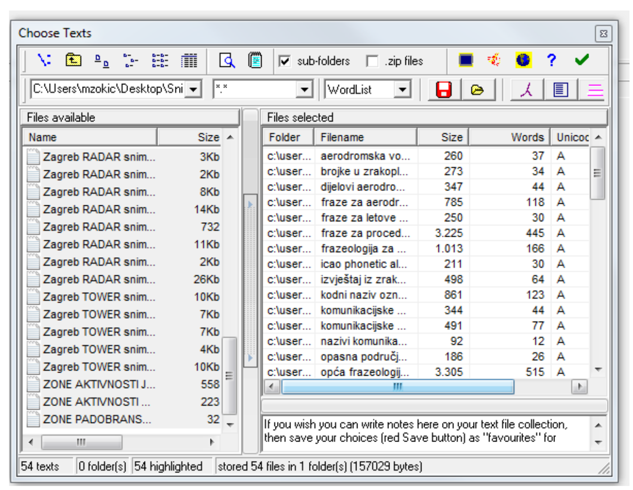


Fig. 6. A list of selected texts.

WordList tools enable us to see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, Concord, gives us a chance to see any word or phrase in context and with key words in a text can be found with KeyWords tools [9].

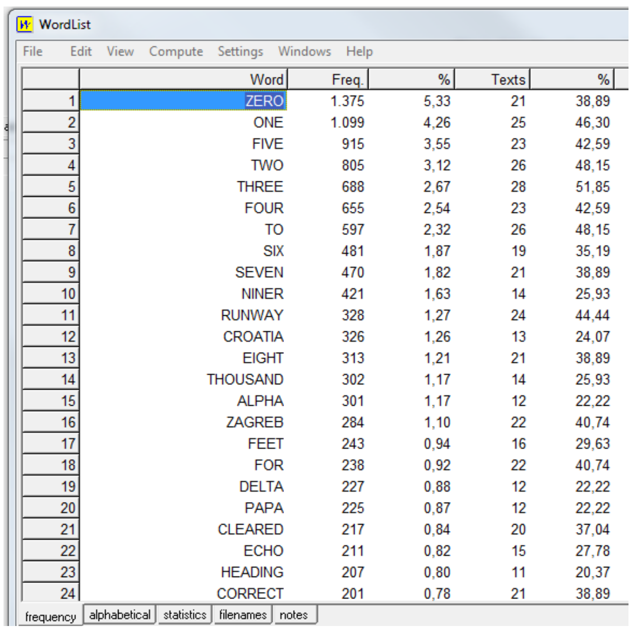
The first step in compiling the corpus was creation of a word list. The WordList tool generates a list of all the words (tokens) or word forms that are included in the compiled corpus and statistical data.

It shows how often each word occurs in the text files, what is the percentage of the running words in the text, and how many text files each word was found in. The words can be listed in alphabetical order and according to the frequency (the most frequent coming first, descending to the least frequent).

54 text files that contain all previously mentioned data were selected to be included in this spoken corpus of radiotelephony phraseology.

The corpus consists of 25828 words (tokens) and 1733 distinct words (types). Type/token radio is 6,75 and mean word length is 4,91. As it can be seen in Figure 8, according to the frequency, the first ten places in the corpus are mostly reserved for numbers. The most frequent word in the corpus is zero. It appears 1375 times. The next one is the word one. The first most frequent lexical word, besides numbers and prepositions, is runway.

The KeyWord tool is a program for identifying the "key" words in one or more texts. Key words are those whose frequency is unusually high in comparison with some norm (some larger corpus; for example British National Corpus).


WordList
File Edit View Compute Settings Windows Help

	Word	Freq.	%	Texts	%
1	ZERO	1.375	5,33	21	38,89
2	ONE	1.099	4,26	25	46,30
3	FIVE	915	3,55	23	42,59
4	TWO	805	3,12	26	48,15
5	THREE	688	2,67	28	51,85
6	FOUR	655	2,54	23	42,59
7	TO	597	2,32	26	48,15
8	SIX	481	1,87	19	35,19
9	SEVEN	470	1,82	21	38,89
10	NINER	421	1,63	14	25,93
11	RUNWAY	328	1,27	24	44,44
12	CROATIA	326	1,26	13	24,07
13	EIGHT	313	1,21	21	38,89
14	THOUSAND	302	1,17	14	25,93
15	ALPHA	301	1,17	12	22,22
16	ZAGREB	284	1,10	22	40,74
17	FEET	243	0,94	16	29,63
18	FOR	238	0,92	22	40,74
19	DELTA	227	0,88	12	22,22
20	PAPA	225	0,87	12	22,22
21	CLEARED	217	0,84	20	37,04
22	ECHO	211	0,82	15	27,78
23	HEADING	207	0,80	11	20,37
24	CORRECT	201	0,78	21	38,89

frequency | alphabetical | statistics | filenames | notes

Fig. 7. A frequency listing for spoken corpus of radiotelephony phraseology

The program compares two pre-existing word lists, which are created using the WordList tool. One of these is a large word list which will act as a reference file. The other is the word list based on one text which is studied. The list of key words has not been created as it is not relevant for this research.

The Concord tool enables us to see lots of examples of a word or phrase in their contexts. This tool is the most important part of WordSmith tools for our research and has most frequently been used in designing the mentioned language technology system. The Concord tool has been used to make a list of phrases that differ from the standard radiotelephony phrase, but have similar meaning and are frequently used and overall accepted. The starting point have been phrases contained in Radio Communication Procedures (a search word or word phrase is specified). Then, the Concord tool looks for it in all chosen text files. And finally, the word or phrase is presented on a concordance display giving access to information about collocates and stored for further usage.

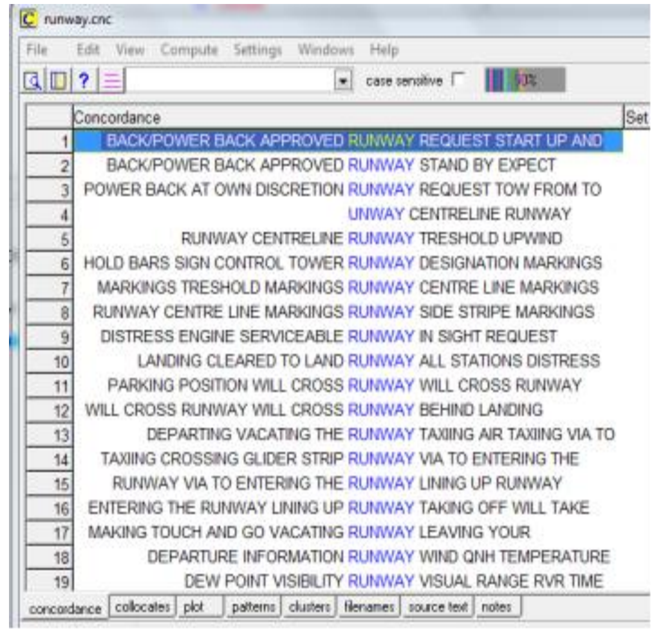


Fig. 8. A list of concordances for the word *runway.*

The concordances can be listed alphabetically or in the order they appear in text files. The listings can be saved for later use, edited, printed, copied to your word-processor, or saved as text files. Figure 8 shows concordances for the word *runway* and its immediate contexts. For the word runway, concordances are listed according to their appearance in the text files.

D. Features of radiotelephony communication spoken corpus

Research in speech recognition, speaker and language identification require the use of corpora whose records contain the variability of the universe of the speakers. Some of the main factors of this variability are gender, age, dialect, recording scenario and others. This implies a high volume of data since all these factors must be well represented. On the other hand, speech recognition experiments cannot be carried out unless information about segmentation and labelling is available [5].

Ideally it is desirable to obtain a large and representative sample of general language. The reason for a large sample is that it could be expected a larger quantity of words as longer is the corpus. This quantity of words will imply bigger dictionary language coverage and it mainly implies greater evidence of the diverse linguistic phenomena required. To be representative supposes several cultural language levels, several themes and genres. However these qualities do not imply each other, instead in some cases they are contrary. One contraposition that must be considered is that between quality and quantity. A big corpus does not guarantee to possess the expected quality. The corpus should be balanced among those qualities [18].

Sometimes it is even more convenient to use a relatively small corpus because the concordances of usage of function words may occupy thousands of pages and most of the examples will be trivial [18]. However, there should be enough texts to reflect relevant features of the dedicated field. The upper limit was connected only with pragmatic considerations, the disk space and the speed of the service software [18].

Here are some features of the designed spoken corpus of radiotelephony phraseology:

a) It is representative. The criterion of representativeness is fulfilled by selection of the text. The corpus contains texts of the same register and content, that is text of radiotelephony communication. The findings from the corpus are generalisable and applicable to European radiotelephony language.

b) In terms of the content, it contains standard radiotelephony phrases prescribed by Radio Communication Procedures, transcripts of radiotelephony communication recordings and terminology used at airports, information relevant for Croatian airspace and information on procedures that are carried out at Zagreb airport.

c) For the moment it can be described as a static corpus. We are aware that this feature may have an influence on the corpus representativeness so the plan is to extend the spoken corpus of radiotelephony communication for future research.

VII. Conclusion

Although the compiled spoken corpus of radiotelephony communication has been designed for Croatian airspace, due to variety of nationalities using Croatian airspace, the designed corpus is found to be representative for radiotelephony language used in Europe and applicable to any research in European radiotelephony communications.

Even though many experts in corpus linguistics agree the larger the corpus the better, for the purpose of this research, a relatively small corpus of spoken radiotelephony language has been designed with only 1733 words. There are two reasons for that:

1. As already mentioned, the language of radiotelephony communication is a restricted, coded and standardized sublanguage with reduced vocabulary.

2. The process of collecting materials for spoken corpus design (recording and transcription of communication) is time consuming.

Nevertheless, it has to be emphasised that smaller specialized corpora containing texts of a particular genre can be extremely useful. It is possible to get much useful data from a small corpus, particularly when investigating high frequency items, as is the case with this spoken corpus. In such corpora is easier to identify specialized terms and detect collocations, and it provides a wealth of information about structure, style and concepts in the specialized target language. All that makes concordancing more representative and utilizable.

References

[1] *AIC – Voice Communication Procedures*, Croatia Control Ltd. Zagreb, 2008.

[2] D. M. Akbar Hussain, M. Z. Khan, Z. Ahmed, K. Ahmad, S. Majeed, M. A. Malik, *A Framework Model through Data Flow Diagrams to Model an Air Traffic Control System*. 6th WSEAS Int. Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, Tenerife, Spain, December 14-16, 2007, p 231-236.

[3] *Communication Procedures including those with PANS, ICAO, Annex 10, Volume 2, 2001.*

[4] E. G. Gallardo, J. O. Hernández, J. A. Segovia De Los Ríos, *Knowledge Representation of Acquisition and Control Systems with Graphical Programming using UML Notation*, WSEAS, ASCOM, 2004, 485-425.

[5] F. Diaz, M. Rubio, P. Gomez, V. Nieto, V. Rodellar, *Using Hidden Markov Models in segmentation of speaker-independent connected-digits corpus*, WSEAS, Shiatos, 2002, 447-259.

[6] G. Hunter, M. Huckvale, *Studies in the Statistical Modelling of Dialogue Turn Pairs in the British National Corpus*, WSEAS, Spain, 2002, 452-187.

[7] G. Sidorov, *Dynamic Management of Text Corpora*, WSEAS, Athens, 2004, 487-818.

[8] J. M. Carroll, *Making Use: scenario-based design of human-computer interaction*. Cambridge: MIT Press, 2000.

[9] *Communication Procedures including those with PANS*, ICAO, Annex 10, Volume 2, 2001.

[10] *Context-Sensitive Speech Recognition in the Air Traffic Control Simulation*, Eurocontrol. EEC Note No. 02/2001. Available: http://137.193.200.177/ediss/schaefer-dirk/inhalt.pdf.

[11] *Effective Pilot / Controller Communications*, Flight Operations Briefing Notes, Human Performance, September 2004. Available: http://www.airbus.com/fileadmin/media_gallery/files/safety_library_items/AirbusSafetyLib_-FLT_OPS-HUM_PER-SEQ04.pdf.

[12] R. Rivera-Lopez, E. Rivera-Lopez, A. Rodriguez-Leon, *Another Approach for the Teaching of the Foundations of Programming using UML and Java*, Proceedings of the 3rd WSEAS International Conference on COMPUTER ENGINEERING and APPLICATIONS (CEA'09), p 279-283.

[13] C. C. Freudenrich, *How Air Traffic Control Works.* Available: http://science.howstuffworks.com/transport/flight/modern/air-traffic-control1.htm.

[14] D. McMillan, *Miscommunications in Air Traffic Control*, Queensland University of Technology, October 1998. Available: http://www.scribd.com/doc/19647051/Miscommunications-in-Air-Traffic-Control.

[15] Pilot-Controller Communications, Eurocontrol. June 2004. Available: http://www.skybrary.aero/bookshelf/content/bookDetails.php?bookId=139.

[16] V. O. Prinzo, *An Analysis of Voice Communication in a Simulated Approach Control Environment*, Civil Aeromedical Institute Federal Aviation Administration, May 1998. Available: http://www.hf.faa.gov/docs/508/docs/cami/9817.pdf.

[17] M. Scott, *WordSmith Tools Manual, Version 6.0.*, Liverpool, Lexical Analysis Software Ltd., 2013.

[18] S. N. Galicia-Haro, A. F. Gelbukh, I. A. Bolshakov, *Compilation of a Mexican Spanish text corpora*, WSEAS, Mexico, 2002, 170.