Transcription and annotation of an apprenticeship corpus: application to the correction and self-correction strategies

Jean-Léon Bouraoui [1], Gwenael Bothorel [2], Nadine Vigouroux [3]

[1, 3] IRIT/DIAMANT
University of Toulouse (Paul Sabatier)
{bouraoui, vigourou}@irit.fr

[2] CENA/PII
ENAC (Toulouse)
bothorel@cena.fr

Abstract

This paper presents a corpus-based study devoted to corrections, self-corrections strategies to recover errors during dialogues. The corpus used results from exercises where air-traffic controllers being formed interacts with people simulating pilots in practice. Considering correction strategies as an answer to errors, we present a fine-grained typology of these strategies, the errors they aims to compensate, and their markers. We also give various statistics about their distribution in the corpus, and comment them, in regards with their application to spoken dialog systems.

Introduction

The study and the conception of human-machine dialogue systems undeniably come through the collection and the analysis of dialogue corpora, whether these ones are obtained from interaction of human to human, human to machine, or else during a Wizard-of-Oz session. The work that this article presents relies on the study of a corpus made of training dialogues between apprentice controllers and persons playing the role of pilots (named "pseudo-pilots").

In the context of controllers' activity, error handling is a very important thing, since it concerns the management of traffic and the security of planes. We study here the way this handling is made. It goes through strategies of correction and self-correction, which are peculiar features of spontaneous speech (since writing normally does not let any trace of correction), especially in stress and apprenticeship situation, as is the case with air controllers in formation. Indeed, because of the necessity of managing errors, each one has imperatively to be detected and corrected as soon as possible. Due to these domain constraints, the aim of this paper is to examine categories of errors through different correction strategies that show among others the speaker behavior facing to error situations and the role of interlocutor.
There is "correction" when one of the participants of dialog notices that he has not been understood by his interlocutor, and correct him by giving the utterance to understand. On the other hand, one speaks about "self-correction" when a speaker realizes himself that the utterance he has just produced (or one of his previous utterances) comprises a mistake, and seeks to rectify it. This distinction is essential for the design of error-recovery dialog strategies (such as rephrasing, repeating, etc. [1]) within an automatic system. In the first case, it is a matter of being able to handle the way the human interlocutor noticed to the system that it made an error, and its nature. In the second case, the system has to understand that the user is correcting himself (and not the system), and what is corrected. The results of such a study also have to be formed in order to be modeled in any automated interaction system. For instance, the system will benefit to determine if the user corrects it or itself, what is concerned by the correction, and where the corrected element is located.
Finally, the necessity to spot out corrections entails the use of specific markers, that is clues that indicates the proximity of a correction. Their knowledge is very important, for example when one seeks to adopt a specific strategy when this phenomenon appears. Thus, the designer of a spoken dialogue system could wish that it will be able to recognize that the user made a self correction, which would entails for the system to correct itself its understanding of the user's request.
In order to deeply study errors and correction and self-correction strategies, we propose here to categorize them. We do the same for their markers. This classification will be supported with statistics on the occurrences of these various categories, obtained from the corpus mentioned below.

In a first part, we will present with much more details the corpus, the context in which it has been recorded, along with the way it has been transcribed (detail which has importance). Then, a precise classification of corrections, errors, and their markers, will be made. Finally, we will display the results obtained, and comments them with regards to their application to spoken dialogue systems.

1. Description of corpus

The corpus on which focus our study is made up of the recordings of spoken and spontaneous dialogues between air-traffic controllers being formed and "pseudo-pilots" (that is, people simulating pilots in practice). Two languages were used: French and English (French being the majority). These dialogues occurred during exercises at the ENAC (Ecole Nationale d'Aviation Civile; in English: National School of Civil Aviation) from Toulouse. We obtained it in the framework of a more general work consisting of transcribing and annotating the recordings.

1.1. Main characteristics of controllers – pseudo-pilots interactions

The aim of the exercises recorded was to train apprentice controllers, and then evaluate them. It consists of managing several planes that are in a controlled area, for example by

assigning them a given speed and/or position. The conditions of exercises were as near as possible from real environment: controllers worked with screen giving the radar position of virtual "planes"; the air traffic was simulated by several persons assuming the role of one or many pilots. The typical situation of communication between a controller and a pilot is represented in figure 1, which also shows the main factors that can decrease the work performances: apprenticeship and workload.

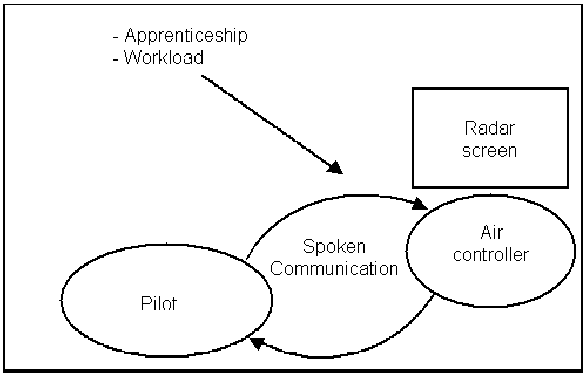

Figure 1: Situation of communication of air traffic control

The utterances produced by the controller, as well as the pilots' ones, are driven by a quite strict phraseology [2]. It describes, for example, the way the speaker must pronounce the planes call signs, or the order that the different components of a message have to follow; also, two speakers can't speak at the same time (it is due too to technical reasons: the communication canal cannot be engaged by more than one speaker). Actually, this phraseology is not always strictly respected, even if its general guidelines are kept. As we will see, some of the correction and self-correction occurrences are relevant to the respect of the phraseology.

A typical instance of a simple order that an air controller can give to a pilot is : "D T C climb level 9 0": we see, first, the call sign of the pilot's plane ("D T C"), and then the order itself. If we want to formalize this, we can say that a prototypal order is composed by a call sign and the order, this one being composed of a command that plays the role of a predicate, whose argument is a value (for instance, a position, like "9 0 " in our example). Some more complex utterances can also occur, composed of a sequence of simple orders. For more details on these topics, see [3].

1.2. Transcription and annotation methods

Our work consisted to transcribe dialogues as well as to annotate them according to some specifications ([4] and [5] respectively for transcription of recordings, and for their annotation). The whole transcription was done by one of the authors; some difficulties were resolved with help of a phraseology expert.

The aim of the work was to reveal various linguistic phenomena. The knowledge that will be obtained could serve to the design of efficient artificial agents that could substitute to the "pseudo-pilots" played by human until now.

The tool we used for transcription is Transcriber; it's a software initially developed at the DGA (Délégation Générale pour l'Armement: in English: General Delegation for Armament) to permit the transcription of broadcast news

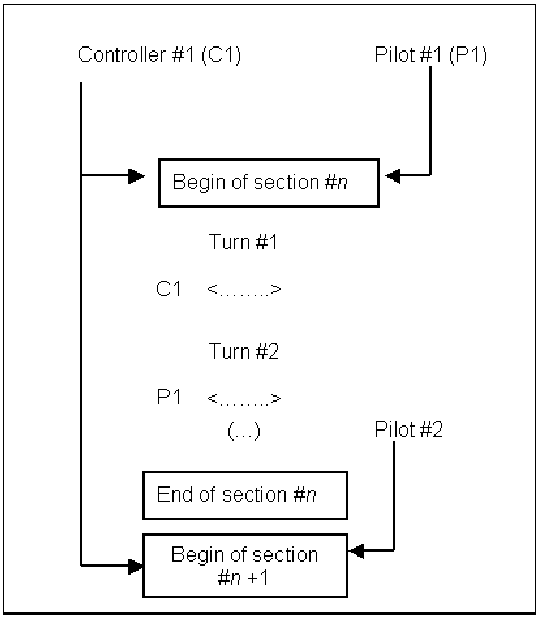

Figure 2: Sequence of sections and turns

([6]). It offers advanced functions of transcription and annotation, one of the most precious being easiness of segmenting dialogue into several structures: sections and turns. We refer by this last term to an utterance produced by a given speaker without any interruption (it occurs when the utterance is complete, or when it is interrupted before end for any reason). A section is a set of turns of speech between a controller and a given pilot, until one (or the two) of the participants of dialogue is substituted by another. The figure 2 above shows the formalization of the sequence of turns and sections.

Furthermore, Transcriber gives opportunity to save transcription under several electronic formats, among which the handiest is XML, conceived to be easily portable and handled. We also made use of the possibility of using a certain number of predefined marks (to mark for example the various noises that can occurs) and to create our own labels, which permits to easily locate the strategies we study in the present article, and the context where they appear. These marks were used for the two levels of annotation consisting our work: transcription of what is said along with the various phenomena that can occur during spoken dialogue (pause, noise, accentuation, etc.) and annotation of strategies employed (notably self-corrections).

Here is an example of an utterance thus marked: " Poitiers D er *[...] [mic] [self_correction]* ENAC D I K C C er good morning level 1 1 0 direct Amboise", where the tags in italics correspond respectively to a pause (see below for a complete definition), a noise from microphone, and a self-correction. We also used the colon ":", useless for its classic role of punctuation mark, it worked as a diacritic to indicate the words pronounced with a particular emphasis (for

instance: "*Paris:*" shows that this name has been accentuated).

These possibilities allow us to simplify statistical enquires, such as counting the number of occurrences of the various strategies.

1.3. Description of corpus

The recordings were made with a DAT (Digital Audio Tape), and sampled at 16 kHz (16 bits). For recording reasons, the sound quality sometimes suffers from problems resulting from saturation or noises such as interferences; however, the speech signal is intelligible.
We present the main features of the corpus in table 1 below.

Table 1: Length and number of speech turns of the recordings

	Length	Number of speech turns	Number of apprentice controllers
Total	19h04mn23s	5946	32 (distributed in 2 groups)

2. Typology of phenomena linked to errors

2.1. Errors typology

After studying the corpus, we noted that, whatever the error is, it's not the whole utterance (simple or complex, as defined in 1.1 above) that is wrong, but only a part of it, or the way it is constructed. In order to take into account this observation, we defined the following classes of errors:

- **Error on a value**: we mean by "value" an alphanumeric data that can be considered as an argument of a command (cf. 1.1); it can be for example a plane call sign ("Britair 452"), a position ("9 0"), a town ("Paris"), etc;
- **Error on a command**: a term (most often corresponding to an order, such as "climb", "request", etc.) is substituted to another;
- **Error on organization**: a word or group of words is not at its correct position in the utterance (for example "Air France 41 82 good morning *climb level* identify climb level 140": here, the speaker realized that he began to give the order "climb level 140" before the order "identify"; consequently, he correct himself). This kind of errors can occur even in a daily dialogue, but is increased by the phraseology, which sets a given order for the components of the utterance;
- **Error on the language used**: the speaker notice (or is being noticed) that he does not speak in the correct language (French instead of English or vice versa; for example, in the following dialog, the pseudo-pilot reminds to the controller that he must talk to him in English: Controller: "N 9 O O euh F R contact ENAC 123 décimale 8 ." – Pseudo-Pilot: "in English please.";). This category is totally dependant of the domain (air traffic control). Indeed, it is due to the fact that the controller has to speak one language according to the pilot he addresses to: this kind of situation can hardly occur in most of everyday tasks!

When an error is noticed, whether it is by the speaker or his interlocutor, it gives rise to various strategies of correction and self-correction, which we describe below.

2.2. Correction and self-correction strategies

We'll make a distinction between three main strategies of correction: self-correction of an element of the utterance being produced, self-correction of a previous utterance, or correction coming from the interlocutor. The distinctive features of these categories are based on the person who does the correction (speaker or interlocutor) and the moment when it occurs. Indeed, we can think that these different kinds of corrections can occur in distinct ways, and consequently be characterized by specific markers. Some studies on others oral corpora [7] also revealed the existence of a phenomenon called "false-start" It occurs when the speaker begins a word, and stops producing it before the end. Of course, we considered it like an other category of self-correction.
Here are examples of each of these categories, taken from our corpus (we set the element being corrected in italics):

- **Self-correction**: "KLM er 2 1 5 climb level 1 9 0 contact ENAC 120 contact ENAC er *1 2 6 decimal 8 5*.". The controller asks to pilot to go to level 190, and to contact ENAC on frequency 126.85; he makes a correction on the frequency to use;
- **Correction of a previous utterance**: here is a short dialog between a controller and a pseudo-pilot: Controller: "er F K C correction maintain level 1 9 0."- Pseudo-Pilot:"to level 1 7 0 K C ." - Controller: "er F K C correction maintain level *1 9 0*." The controller first gives a position to which the pseudo-pilot must go; this one confirms, but afterward, the controller corrects his previous order, that was giving wrong coordinates;
- **False-start**: "F G H M N ENAC good morning (…) speed minim er 200 Knots *minimum*.". The speaker begins to utter the word "minimum", and stops himself before ending it for he noticed that he had not gave the speed;
- **Correction from the interlocutor**: here again, a dialog between a controller and a pseudo-pilot: Controller: "euh TAT 289 M L (…) join Poitiers" - Pseudo-Pilot: "Lacan Amboise Poitiers it's TAT M *I*.". In this example, the controller made a mistake on a part of the call sign of his interlocutor; consequently, this one corrects him.

2.3. Markers

This part will be subdivided in two: we will first make general remarks about the different markers picked out, and then focus on the case of lexical ones, which present some interesting features that we display in section 2.3.1.

2.3.1. *General remarks*

Two questions rise when one speaks about makers of a given phenomenon: what is the length of the scope around the phenomenon where something can be considered as marker, and which are the kinds of markers searched. Here are the answers we provided to these questions after observing the corpus:

- We fixed the scope to 3 words before and after the correction phenomenon itself; this value results from empirical observations, as well as from the fact that some three "words" sequences form in fact one unit (as for call signs for example; for more details on that point, see [3]);
- Three different kinds of markers were used: lexical, accentual and finally others kinds of spontaneous speech phenomena. The two last ones results from the oral nature of the corpus: we employ the term "accentual" to designate the emphasis put on a word by the means of a variation of prosody. Thus, when a speaker corrects a wrong element within a call sign, it arrives that the element being corrected is pronounced with a particular accent. For example, in "Lacan Amboise Poitiers it's the TAT M *I*" (previously mentioned), the element in italics, that corrects a wrong value previously given, has been accentuated by the speaker, The global designation "spontaneous speech phenomena" puts together various phenomena such as hesitations, pauses (we call pause a non-speech period during more than half a second: we formulated the hypothesis that a silence during such a length is revealing of an enunciation problem such as the thought time necessary to find the correct word to say) or repetitions (contrary to [1], we didn't put them in a specific category).

2.3.2. *Lexical Markers*

Among the lexical markers, we made the following distinction, from what we observed:

- **Deictic**: word referencing to other word, such as "it's" (or "c'est" in French). The most frequent configuration is the following: "it's CS" (where CS is a call sign; for instance: "it's A M L 753"). One should note that this usage of deictics also occurs quite frequently used in other contexts, especially by pilots to introduce themselves;
- **Excuse**: for example, "sorry", "excuse me", etc.;
- **Negation**: any words used in order to negate something, the most common one being "no";
- **Correction**: the word "correction". We put it in specific category, for its usage is explicitly asked by the phraseology for marking the correction of an utterance; according to it, the correction must then be followed by the element corrected.

3. Results and comments

We'll display our statistics according to the classification presented in section 2 above: firstly errors, then correction and self-correction strategies, to conclude with their markers.

3.1. Errors

On table 3, the reader will find the number of occurrences and the percentage (calculated in comparison with the total number of errors) of each category.

Table 2: Number and percentage of errors categories

	Number	Percentage
On a value	81	1,36%
On a command	52	0,87%
On organization	12	0,20%
On language used	6	0,10%
Total	**151**	**2,54%**

There's the same number of noticed errors that of corrections. This is normal: as said in introduction in air traffic control, any error must be corrected at a moment or another, the sooner being the best. Most of the errors concerns what we called "values", along with "commands". It is not surprising since nearly all utterances contain at least one reference to a call sign, a speed, etc. The same reasoning can be applied to "commands". However, there is 1.5 times less errors committed on "commands" than on "values". According to us, this can be explained by the fact that what we called "values", especially call signs and positions are quite complex sequences of numbers and letters, used only in air traffic control context. Consequently, they certainly require handling an important cognitive load, thus leading to more errors. We also think that this can explain the lesser number of errors of command utterances (nearly two times less occurrences than for "values") and of organization (more thansix times less occurrences than for "values").

3.2. Corrections and self-corrections

In table 2, we display the number of occurrences of the different kinds of correction found in the corpus, along with their percentage in comparison with the number of speech turns. This last result must be tempered, since there are sometimes several corrections occurrences for one speech turn; in spite of this, it gives a good idea of the global proportion of this phenomenon through the corpus.

Table 3: Number and percentage of corrections strategies

	Number	Percentage
Self-Correction	109	1,83%
Self-Correction of a previous utterance	10	0,17%
False-start	27	0,45%
Correction by interlocutor	5	0,08%
Total	**151**	**2,54%**

It appears that the most frequent kind of correction is the first one (the speaker corrects himself, during his current utterance). It is difficult to draw any other conclusions, since this would imply to have at one's disposal corpora from others domains, determining the number of utterance it contains, and compare the two results.

3.3. Markers

The results are given in number of occurrences and percentage (calculated in comparison with the total sum of markers) in table 4.

Table 4: Number and percentage of markers

	Number	Percentage
Lexical	25	17,99%
Accentual	16	11,51%
Spontaneous speech	98	70,50%
Total	**139**	**100,00%**

A few remarks have to been made before commenting these results. First, that these markers, especially accentual and speech spontaneous ones, occur quite frequently in the corpus, and not in the one context of corrections: consequently the results should be handled with care; however, we make the hypothesis that, when located in the context of a correction phenomenon, they are directly linked to it, and hence can be considered as its marker. One can also take into account the fact that there can be several different markers for a single correction occurrence; but it does not affect the global distribution given in table 4.

The first obvious thing is that there are less markers than correction occurrences, despite the fact that, as we said there can be many markers for a correction. This is due to the fact that it is possible to find errors and/or corrections even in lack of markers. This can seem odds; in fact, it is because of the nature of corpus: both phraseology and limited vocabulary put such restrictions on utterance production, that practice allows to recognize almost any error.

We also found useful to consider these in regard to the categories of errors and of correction. The results are displayed respectively in tables 5 and 6 (number of occurrences).

Table 5: Distribution of markers according to correction categories

	Self-Correction	Self-Correction of a previous utterance	False-start	Correction by interlocutor
Deictic	0	0	2	0
Excuse	5	0	0	1
Negation	5	0	0	0
orrection	3	9	0	0
Total	**13**	**9**	**2**	**1**

The most noteworthy result concerns the "self-correction of a previous utterance" category. The fact that almost all of its markers belongs to lexical kind shows the respect of phraseology, which stipulates that the correction of wrong previous utterance has to be indicated by the word "correction". It appears more clearly when one considers the "self-correction" category: all kinds of markers can appear in its context, the most frequent being those belonging to spontaneous speech. An explanation could be that, since the correction occurs during the current utterance, the speaker reacts immediately, without thinking to the phraseology.

Table 6: Distribution of markers according to errors categories

	On a value	On a command	On organization	On language used
Lexical	20	3	0	2
Accentual	11	4	1	0
Spontaneous Speech	58	27	10	3
Total	**89**	**34**	**11**	**5**

The four kinds of lexical markers are arranged as follows (the percentage is calculated in comparison with the total number of occurrences):

Table 7: Number and percentage of lexical markers.

	Occurrences	Percentage
Deictic	1	4,17%
Excuse	5	20,83%
Negation	4	16,67%
Correction	14	58,33%
Total	**24**	**100%**

We see that the most frequent lexical marker is undoubtedly the word "correction", which of course can be attributed to the respect of phraseology guidelines.

Tableau 8: Distribution of lexical markers according to correction categories

	Self-Correction	Self-Correction of a previous utterance	False-start	Correction by interlocutor
Lexical	13	9	1	2
Accentual	13	1	1	1
Spontaneous speech	78	0	20	0
Total	**104**	**10**	**22**	**3**

Two phenomena have to be noticed. First, that once again it is in the self-correction category that the various kinds of

lexical markers are the most equally arranged. Second, that we note the opposite phenomenon concerning self-corrections of a previous utterance: all occurrences are marked by the word "correction", which can be imputed directly to respect of phraseology, as for table 5. The arrangement related to false-starts and correction are far too few to be actually interpretable.

Tableau 9: Distribution of lexical markers according to errors categories

	On a value	On a command	On organization	On language used
Deictic	0	0	2	0
Excuse	5	0	0	1
Negation	5	0	0	0
Correction	3	9	0	0
Total	**13**	**9**	**2**	**1**

It is interesting to see that most of the occurrences of the word "correction" have been made for correcting a command (and not a value). Since, as we saw in table 8, that this marker occurs only for self-correcting a previous utterance, we can conclude that what we called "value" are more likely to be immediately corrected in current utterance than commands. The reasons of this phenomenon should probably deserve a cognitive analysis.

4. Conclusions

We have studied a corpus of spontaneous speech dialogues, consisting of interactions between air controllers and pseudo-pilots. We put focus on errors, corrections strategies, as well as their markers, notably by arranging them among within various categories.

We saw that the most frequent kinds of errors concerns what we called "values", such as call signs. In terms of application, that could means that a spoken dialogue system in a learning environment should have increased robustness errors on these elements of utterance. Concerning the markers, we observed a large amount of lexical ones, as well as spontaneous speech phenomena. We established a link between the use of lexical markers and the phraseology that directs the production of utterances.

More generally, the phraseology plays an important role for some of the errors that occur. For example, it is the case when the cause is a deviation regarding to the organization of the utterance, which, as we said, is less primordial in daily life. It also affects our interpretation of corpus, notably for detection of errors even without markers. It reduces the vocabulary and the number of syntactic patterns, thus allowing, for example, an easier implementation understanding component. But it also has a negative effect: the results we obtained are strongly linked to the one domain of air control interaction This lead us to the wish that this study could be extended on others corpus pertaining to various domains, which could notably permit a deep evaluation of actual phraseology on the way errors and corrections are made, and maybe to develop a general and common nomenclature of those strategies. Some new studies in that way, such as [8], begin to appears, even if they are not always grounded on the same basis and aims [1].

5. Acknowledgements

This study is funded by the CENA.
We thanks Philippe Truillet for the records of the exercises, and the information he gave us about it.

6. References

[1] J. Shin, S. Narayanan, L. Gerber, A. Kazemzadeh, D. Byrd (2002) "Analysis of user behaviour under error conditions in spoken dialogs", *Proceedings of the Seventh International Conference on Spoken Language Processing*, Colorado, September, 2002.

[2] *Arrêté du 27 juin 2000 relatif aux procédures de radiotéléphonie à l'usage de la circulation aérienne générale*, J.O n° 171 du 26 juillet 2000, p. 11501.

[3] Dourmap, L,. Truillet,T. "Interaction vocale dans le contrôle aérien : reconnaissance automatique d'indicatifs d'avion"*, CENA internal report*, 2003.

[4] Coullon I., Graglia L. "Spécifications de la base de données pour l'analyse des communications VHF en route", *CENA internal report*, 2000.

[5] Coullon I., Graglia L., Kahn J., Pavet D. "Définition détaillée du document type (DTD) pour le codage sous XML des communications VHF en route – VOCALISE Trafic CRNA / France 2000", *CENA internal report*, 2001.

[6] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman "Transcriber: development and use of a tool for assisting speech corpora production", *Speech Communication special issue on Speech Annotation and Corpus Tools*, Vol 33, No 1-2, 2000.

[7] Bousquet, C. *Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique*, PhD Thesis, University of Paul Sabatier, Toulouse, 2002.

[8] Bousquet-Vernhettes, C., Privat, R., Vigouroux, N. "Error handling in spoken dialogue systems: toward corrective dialogue", *ISCA workshop on Error handling in dialogue systems*, 2003