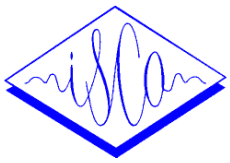

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 33-37.

Disfluency phenomena in an apprenticeship corpus

Jean-Leon Bouraoui & Nadine Vigouroux

[*]IRIT, Toulouse, France

Abstract

This papers presents a study carried out on an apprenticeship corpus. It features dialogues between air traffic controllers in formation and "pseudo-pilots". "Pseudo-pilots" are people (often instructors) that simulate the behavior of real pilots, in real situations.

Its main specificities are the apprenticeship characteristic, and the fact that the production is subordinate to a particular phraseology.

Our study is related to the many kinds of disfluency phenomena that occur in this specific corpus. We define 6 main categories of these phenomena, and take position in regard to the terminology used in literature. We then present the distribution of these categories. It appears that some of the occurrences frequencies largely differs from those observed in other studies. Our explanation is based on the corpus specificity: in reason of their responsibilities, both controllers and pseudo-pilots have to be especially careful to the mistakes they could do, since they could lead to some dramas.

The remainder of our paper is dedicated to the more deepen study of a disfluency class: the "false starts". It consists of the beginning utterance of a word, that is not achieved. We show that this category consists of several sub-categories, of which we study the distribution.

1. Introduction

It's beyond doubt that disfluencies do occur very frequently in everyday conversations. Many studies are devoted to these phenomena, and to their various manifestations. The majority of this studies is carried out on corpus made of everyday life productions.

But one can wonder what would be the manifestations of disfluencies in a corpus made of very specialized and constraint language. This is a very important question, since answering to it would give us very important hints on disfluencies. If one find that they occur in the same way than in everyday language, that proves that some "universal" classes of disfluencies do exist. If it didn't, then it would be very interesting to study the nature of their differences, and their cause. In any case, this entails some additional knowledge on the disfluencies. Beyond the single evident theoretical interest, it also offers a wide range of practical applications, notably in automatic speech recognition and understanding.

The work we present in this paper is precisely devoted to a study carried out on a corpus such as we described above. In the first part, we present in details its characteristics. We also explain the methodology we used to transcribe and annotate it. The second part of this article relates to the precise description of the phenomena we sought, and the results we obtained.

2. Description of corpus

2.1. Characteristics of controllers – pseudo-pilots communication

The formation of the Air Traffic Control (ATC) controllers includes theoretical teachings, but also consists of a lot of training sessions. These sessions are made of communication between air-traffic controllers being formed and "pseudo-pilots operators" (that is, people simulating real pilots).

The aim of the exercises is to train apprentice controller activities, and then to 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. Two languages are used: French and English (French being the majority); all the speakers are French native speakers. The exercise conditions are as near as possible from real environment: controllers work with screen giving the radar position of virtual "planes"; the air traffic is simulated by several persons assuming the role of one or many pilots. Some background noises (overlapping conversations, sounds emitted by microphones, etc.) also occur.

The utterances produced by the controller, as well as the pilots' ones, must respect a phraseology [1]. It describes, for example, the way the speaker must pronounce the plane call signs, or the order that the different components of a message/an utterance have to follow. Two speakers can't speak at the same time, due to technical limitations: the audio channel is only assigned to one speaker. During the formation step, the phraseology is not always strictly respected (neither in real work conditions), though its general guidelines are kept. However, its learning and mastering is also aimed by exercises.

An instance of a simple order that an air controller can formulate to a pilot is: "D T C climb level 9 0": we find, first, the call sign of the pilot's plane ("D T C"), and then the order itself. More complex utterances can also occur, composed of a sequence of simple orders. For a complete description of the French call signs and orders, see [5].

The use of the phraseology entails that the lexico-syntactic schemas are limited in number. Moreover, due to the restrictive task, and to the apprenticeship property, it is obvious that, from a linguistic point of view, the corpus differs a lot from a more "traditional" one. We hypothesize that this may probably influences the phenomena that appear, in comparison to some less constraint tasks, such as daily conversations or train timetable reservation for instance.

To conclude, it is important to note that the spoken dialogues are actually spontaneous speech. We insist on that point, because the important role played by the phraseology could make one think that all utterances are already planned. It's not true since both controllers and pilots do not know what will happen, and consequently what has to be said. The phraseology only set up a general framework for utterances; what is actually said depends on the dynamic interaction between a given controller and pseudo-pilot.

2.2. Transcription and annotation methodology

We transcribed dialogues as well as annotated them according to some specifications ([3] and [4]). These authors made a distinction between the orthographic transcription and annotation, which corresponds to an interpretation (at semantic, dialogic levels, etc.) of the orthographical string.

Specifications were defined, firstly to determine elements that has to be transcribed, and secondly to obtain homogeneity of transcriptions in case where several annotators processed the tasks. They consist essentially of rules to follow to transcribe technical ATC items such as call signs, speeds, etc. It also gives instructions to transcribe phenomena such as hesitations, pauses, or accentuations. While transcribing the formation corpus, we believed that this specification wasn't sufficiently fine grained to mark out specific phenomena. Consequently, we contributed to it by creating other classes of phenomena necessary to transcribe, and by refining existing one with sub-categories. Indeed, we considered the fact that the annotator could possibly not have access to the recordings, or not have time to refer to it for a given detail. So, it is necessary to spot any phenomenon that could be interpreted as a marker for a language act, and accessible only via recordings hearing.

It appears that, by doing this, we reach beyond the framework of "raw information" given by specifications, since this decision is based upon an interpretative classifying activity. However, we thought that if it wasn't done during the transcription, the annotator would miss some interesting phenomena.

We used Transcriber 1.4.2 to carry out the transcription.

2.3. Description of corpus

The recordings were made on July 2001 at the ENAC (*Ecole Nationale d'Aviation Civile*; in English: National School of Civil Aviation) from Toulouse.

They were sampled at 16 kHz (16 bits). A DAT (Digital Audio Tape) was used. For recording reasons, the speech signal quality sometimes suffers from saturation or noises such as interferences. However, it stays intelligible.

We present the main features of the corpus in table 1 below.

Table 1: Main characteristics of our corpus.

Length	Number of speakers	Number of "exchanges"	Number of speech turns	Number of words
36h50mn	16 (distributed in 2 groups)	2 019	11 427	76 306

3. Disfluency phenomena

3.1. Terminological considerations

In literature, many different words are used by various authors to refer to a same disfluency phenomena. This is why it is very important to present which terminology we use here, and the phenomena it designates.

We describe here 6 main classes of disfluencies. Whenever it is necessary, we give an example of the phenomenon they correspond to.

- Hesitations: this term only designates the interjection "euh", which corresponds to "er" in English. It is usually considered to point out a moment's thought on what has to be said next. According to some terminologies (notably [7]), it belongs to the category of "filled" pause. Example:
  maintenons niveau 1 0 0 Poitiers Amboise euh Lacan
- Repeated words: we gave a slightly restrictive definition to this one. We called "repeated words" any situation where a word (or a group of words) appears at least two times consecutively. We do not take into consideration any repetition of a disfluency phenomenon, such as hesitations, fragments, etc. Example:
  station station calling euh repeat your callsign
- False starts: one of the word which has the most various meanings according to authors. We use it to refer to the utterance of the word that does not come to an end. It is worthy to note that in our own terminology, a false start always corresponds to a fragment of word that can be identified. The knowledge of the phraseology helps a lot for this identifying task. Let's see for example the following example; we put the false start within brackets:
  speed euh 200 Kts [mak] euh minimum

The context (both of the previous utterances and of the situation) and phraseology help to understand that the speaker first began to utter "maximum". He realized that he was wrong, and stopped the production ("mak"). Finally, he said the correct word: "minimum".

- Fragment: contrary to "false starts", "fragments" refers to a sound (usually a single phoneme) that can't be identified as a part of a word, or that clearly does not belong to any lexicon. We do not include in this category physiological sounds (breathes, cough, etc.). Example (the fragment is within brackets):
  due to [ou] due traffic euh descend level 9 0
- Lengthening: a lengthening occurs when the production of a sound (usually a phoneme) lasts more than usual. According to some authors, it also belongs to "filled pause" category. While transcribing and annotating the corpus, we took as minimum value for lengthening 20 cs (centiseconds), as many authors in literature.
- Long pause: it is pause (a silent one) that lasts more than 20 cs. We only take into account pauses that occur during a given speaker's speech turn (and not, for example, between two speech turns from different speaker).

Now, we see in the next section the distribution of those phenomena, and do some comparisons with results obtained in similar studies. As we'll see, some results differ a lot from the average observed on others corpus.

3.2. Distributions of phenomena

The Figure 1 displays the distributions of the phenomena described in section 3.1. The number in bold corresponds to the total number of occurrences in the corpus; the percentage is computed in regard to this total number. The sector corresponding to the word repeats does not actually appear on the graphic since it is below 1%.

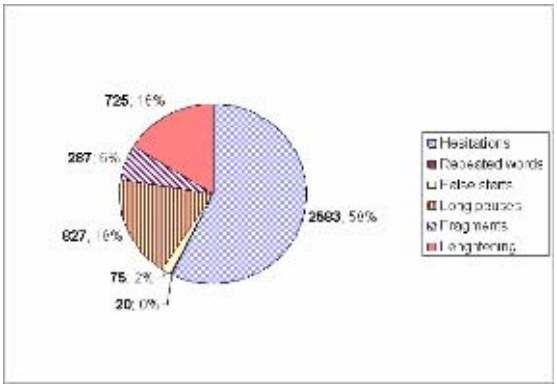

Figure 1: Distributions of the various disfluencies in our corpus.

This distribution calls for many commentaries. In this paper, we will proceed to a detailed comparison with some others studies. Here is a short description (nature of the task, number of words, etc.) of each spontaneous speech corpus on which these studies are based:

- [2]: the work presented in [2] is based on 13 tales orally told by children. It lasts 70 minutes and 25 seconds;
- [6]: based on a 1 000 382 words corpus (various spoken situations; 794 different speakers);
- [7]: based on a corpus lasting 54 minutes, and comprises 8500 words. The 10 different speakers talk about their job or their memories;
- [8]: this thesis is based on a corpus consisting of negotiations (in English language) of merchandise transport by train. It comprises 52 000 words.

As we see, these studies are based on some very different corpus, whether in task or in length. This diversity will thus constitute a valid basis of comparison with our own corpus.

We will now present the comparisons for the categories of disfluencies that we defined. Of course, since any study does not cover all the disfluencies, we will only present those that concerns a given phenomena, or which categorization is close from our. These comparisons would need to be deepen, since in most of the cases, the categories we defined are more or less slightly different from those from other studies. But it will give us a good overview of the specificities of our corpus.

Table 1: comparison for repeated words

Name of the study	Our corpus	[2]	[7]	[8]
Number of repeated words	20	110	141	256

- Repeated words: the most surprising result concerns the number of word repeats, as it appears in table 2.
  Indeed, it appears that the frequency of word repeats are always considerably higher than in our corpus. How to explain such a difference? First, remind that we do not take into account repetitions of any phenomena of disfluencies, even false starts. But that do not explain the huge difference of number. We think that the main explanation is the very nature of the corpus. Our hypothesis is that, in the ATC situation, the speakers (both controller and pilot) can not afford to produce any ambiguity or problem that could affect the comprehension of the utterance. Also, the time necessary to produce an utterance is not extensible: the speaker must not spend too much time in hesitation or other pauses (filled or silent). As we will see next, this hypothesis is confirmed by the fact that for all of the disfluencies that we defined, there is always less occurrences in our corpus (proportionally to the size of the corpus of comparison).

Table 3:

Name of the study	Our corpus	[2]	[8]
Number and/or percentage of hesitations (in regards to the total number of words)	2583 3..38%	554	3512 6.75%

- Hesitation: as we saw in table 3 below, there is also much less hesitations in our ATC corpus than in other studies. There is admittedly 544 occurrences in [2], but this corpus is 70 minutes long, whereas our is 35 hours long. So, there is proportionally more occurrences in the corpus used by [2]. However, one can notice that the difference seems to be overall lesser than what we observed for repeated words.
- False starts and fragments: among the studies we chose to compare with, [6] is the only one whose categorization is the closest from our, resorting to what we call "false starts" and "word fragment". It also present detailed statistics about their distribution. As the authors do not distinguish between "false starts" and "word fragment", we will add up the occurrences of both phenomena that appear in our corpus. The result is a total of 362 occurrences, i.e. 0.47% of the total number of words. [6] reports a total of 6094 occurrences of "word fragments" for about 1 000 000 words (approximately 0.6%). Thus, the distribution in our corpus of this twofold category is quite close of the one observed in [6], contrary to what occurs for the others categories. But this result might be due to the fact that this twofold category do not exactly match with the one defined by [6].

Table 4: Comparison for lengthening

Name of the study	Our corpus	[2]	[7]
Number and/or percentage of lengthening (in regards to the total number of words)	725 0.9%	284	669 (including "euh") 7..9%

- Lengthening: again, as we see in table 4, the frequency of what we called lengthening is lower in our corpus than in the other ones.
- Long pause: the table 5 shows that the specificity of our corpus is a little less pronounced than for the other categories of disfluencies. But, here again, we notice that there is less "long pauses" than in other corpus.

Name of the study	Our corpus	[2]	[7]
Number and/or percentage of long pauses (in regards to the total number of words)	827 1.08%	147 1	318 3.74%

Many pages would be necessary to exhaustively examine the different phenomena, the differences observed with other study, and their causes. In the framework of this paper, we will only focus on false starts. They are the object of the next section.

4. A study on "false starts"

Why focusing on false starts? As we saw, they are far from being the most frequent disfluencies in our corpus. All the same, we think they are worth of interest, for two main reasons. First, they are special cues on the "work of formulation" (we take up here the expression used in [2]). Notably, they can show, in some case, the word that the speaker has in mind. Thus, they help to base some hypothesis on the nature of the problem. Secondly, though in little number, they manifest themselves in different ways that are interesting to identify.

In the first section, we present the different kinds of false starts we found in our corpus. Then, we present the distribution of those categories.

4.1. The different types of false starts

First able, we found out that two kinds of false starts do exists They differ according to their function in the production of the

utterance. The first one do not have any visible function. Example:

route Lacan [amboi] Amboise Balon Limoges

It is probably useless to seek out any function in this category. This kind of false starts is only a mark of "the work of formulation". 29 occurrences of false starts, that is to say 39% of the total number belong to this category.

To the contrary, the second type plays a role of correction of a mistake that was about to be done. Let's see for example the following example:

[mike] Paris 124 decimal 05 Littoral M C

This concerns 46 occurrences, i.e. 61% of the total number of false starts.

This last category can give us some precious hints on the behavior of the speaker and the causes of his errors. Consequently, they deserve a deeper analysis. Many works on disfluencies (for instance, [2], [6], [7], [10]) carry out their study by analyzing the distribution of a given disfluency phenomena according to the lexico-syntactic category of the word it affects. For now, we prefer not to do so. Indeed, our corpus specificities require a specific linguistic characterization. For instance, it is difficult to make a distinction between "function words" VS "lexical words".

In the meantime, we do a typology of the false starts according to the "role" of the word or group of words they affects (except for a category). By "role", we mean the function assumed in regard to the phraseology. We define the following sub-categories:

- Errors on a "word": we quote the term "word" for he designates commands or order (such as "climb", "request", etc.) but also call signs ("Britair 452" for example). Here is a typical example:

  climbing for level 1 7 0 [mak] euh minimum D M C

- Errors on utterance organization: they occur when a word (or words group) does not appear at the position in the utterance that is requested by phraseology. Example:

  [poi] Absie Poitiers Balon Reson Britair B X

  In this example, the speaker begins to say "Poitiers" at the start of the utterance. But he realize that the name of this town must be said after "Absie". This explains why he stops the first production of "Poitiers".
- Errors on the language used: when the speaker talks in an other language than the requested one. Example:

  P I [vite] speed 2 1 0 Kts

Here, the speaker is supposed to speak English. Or, he begin the production of "vitesse" (the French word for "speed"). This is why he stopped before the end of the production.

- Errors of pronunciation: contrary to the previous ones, this category is not linked to a problem in regard to the phraseology. It appears when the speaker use the correct word, at the proper position, but mispronounce it. For instance, in the following example, the speaker mispronounced the word "Littoral":

  It's [lio] Littoral

As we proceeded for the disfluencies, we will now present the distribution of these categories.

4.2. Distribution of repair false starts

The figure 2 below shows the distribution of the categories described above. As in figure 1, the numbers in bold correspond to the number of occurrences, and the percentage is computed in regard to the total number of repairs false starts.

Most of the errors concerns incorrect "words"; the second most frequent category is the incorrect "word" position. We see there a confirmation of one of our main hypothesis: most of the errors are directly linked to the most unusual (in regard to everyday language) sides of the phraseology and of the task. Thus, "words", such as we defined them, are often call signs, i.e. complex sequences of letters and numbers. It is obvious that they are difficult to handle, especially for an apprentice controller. This leads to a high cognitive load, that itself generate some troubles of production. The same reasoning can be applied to errors related to position of "words".

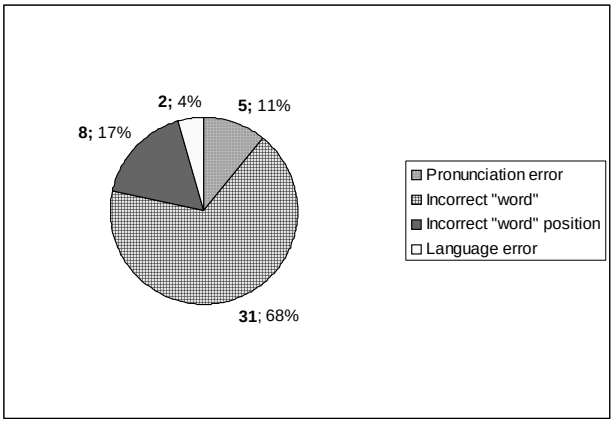

Figure 2: Distributions of the categories of false starts.

5. Conclusion

The corpus on which we lead our study presents many differences with more usual corpora. Indeed, it results from an apprenticeship task. Besides, it requires the use of a phraseology, what entails very restricted lexico-syntactic schemas.

However, it contains many various disfluency phenomena. We showed that there is some differences between the distribution observed in our corpus, and those on other corpus. This is especially true for word repetitions. They seem to be a phenomena that is very sensitive to the task. We explain these differences, on one hand by the phraseology, and on the other hand by the fact that controllers have to be very careful not to produce ambiguities that could have disastrous consequences. We also seen that, more generally, there is less disfluency phenomena in our corpus. We explain this by the same reasoning that we use for word repetitions.

We also studied the specific case of "false starts". They sometime assume the function of corrective instances, or are just some cues for the formulation of utterances. We showed that they can be seen as evidences that the huge cognitive load induced by the apprenticeship is responsible for errors.

Of course, this analysis needs to be deepen. We account to do this from three perspectives. First, we will lead a study of the linguistic properties of the corpus. As we saw, they are very specific, and to compare with others study, it is necessary. We can also take into account the speech rate of

speaker, which is faster than in usual dialogues, and the differences according to the language used.
Of course, we also plan to further deepen the study of the different disfluencies, notably by seeing the correlations between them, as well as with problems that can arise, such as corrections and languages troubles.
The third perspective is to determine the precise nature of the relation between disfluencies and the cognitive load that we attributed to the apprenticeship task. For this, we have at our disposal a corpus made of recordings of dialogues between controllers and pilots that are in real situations. We will apply the same methodology of annotation and study to this corpus. It will permit us to compare the rate of disfluencies in the two different corpus, and to check our hypothesis on apprenticeship influence on their manifestations and frequency.

6. Acknowledgements

This study is funded by the CENA. We thank Philippe Truillet for the records of the exercises, and the information he gave us about it. We are also very grateful to Gwenael Bothorel, who helped us a lot for the transcription/annotation task.

7. References

[1] 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.

[2] Candéa Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. Étude sur un corpus de récits en classe de français.* Ph.D. thesis, Université Paris III (Sorbonne Nouvelle).

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

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

[5] Dourmap Loic & Truillet Philippe. 2003. Interaction vocale dans le contrôle aérien : la comparaison de deux grammaires contextuelles pour la reconnaissance des indicatifs de vol, *CENA internal report*, NR03-669.

[6] Henry Sandrine & Pallaud Bertille. 2004. Word fragments and repeats in spontaneous spoken French, *Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop*, 5-8 Septembre 2003, Göteborg University, Suède, p. 77-80.

[7] Henry Sandrine, Campione Estelle & Véronis Jean. 2004. Répétitions et pauses (silencieuses et remplies) en français spontané, *Actes des XXVèmes Journées d'Etude sur la Parole (JEP'04)*, Fès (Maroc), p. 261-264.

[8] Kurdi M.- Z. 2003. *Contribution à l'analyse du langage oral spontané.* Ph.D. Thesis, Université J. Fourier, Grenoble, France.

[9] Nakatani C. H. & Hirschberg J. (1994), A corpus–based study of repair cues in spontaneous speech, *Journal of Acoustical Society of America*, 953, p. 1603-1661.

[10] Shriberg, Elizabeth. 1994. *Preliminaries to a theory of speech disfluencies.* Ph.D. thesis, University of Berkeley, California.