Etude de dysfluences dans un corpus linguistiquement contraint

Jean-Leon Bouraoui, Nadine Vigouroux

IRIT,
Université Paul Sabatier
118, route de Narbonne,
31062 Toulouse, France
{bouraoui,vigourou}@irit.fr

ABSTRACT

This paper presents a study carried out on an air traffic control corpus which presents some specificity: apprenticeship situation, 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 corpus, and the way they are or not affected by the nature of the corpus. We define 6 main categories of these phenomena. We then present the distribution of these categories. It appears that some of the occurrences frequencies largely differ 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.

1. INTRODUCTION

Les dysfluences sont un phénomène apparaissant fréquemment dans toute production orale spontanée. Elles ont donné lieu à de nombreuses études, que ce soit dans le domaine du Traitement Automatique de la Parole, ou celui du Traitement Automatique du Langage Naturel. En effet, leur étude et leur identification précise est primordiale. Sur le plan théorique, pour mieux comprendre et modéliser les problèmes pouvant advenir lors de toute communication orale. Sur le plan applicatif également, les intérêts sont nombreux : par exemple, pour améliorer la robustesse des systèmes automatiques de dialogue oral, ou « nettoyer » automatiquement des transcriptions de productions orales spontanées (Adda-Decker *et al.* [1]).

Cependant, la majorité des études faites sur le sujet portent plus ou moins sur un langage quotidien : dialogues « à bâtons rompus », demandes d'horaire, discours, etc. Par contre, aucune étude n'a à notre connaissance été menée sur les dysfluences apparaissant dans un corpus d'oral spontané produit dans le cadre d'une tâche particulièrement contrainte, notamment d'un point de vue linguistique, par l'utilisation d'une phraséologie spécifique. Or, on sait que l'utilisation d'une « langue de spécialité »[1] dans le cadre d'une tâche donnée entraîne des comportements spécifiques. Ceux-ci touchent les plans linguistique et cognitif (cf. notamment Lerat [12] et Falzon [8]). On peut alors se demander si on peut également observer ce type de modifications concernant l'apparition des dysfluences dans l'oral spontané.

Répondre à cette question, ou du moins apporter des éléments précis d'information, est précisément le but de l'étude relatée dans le présent article. Elle porte en effet sur un corpus d'oral spontané consistant en dialogues relatifs au contrôle aérien, et devant respecter une stricte phraséologie.

La présentation de notre étude se fait en trois temps. Le premier, assez succinct, est consacré à la description du corpus et de ses caractéristiques. Dans un deuxième temps, nous donnons les différents types de dysfluences relevées dans le corpus, ainsi que leur distribution. Enfin, nous nous intéressons à un phénomène particulier de dysfluence, et à la manière dont celui-ci est affecté par la nature du dialogue et de la tâche.

2. PRÉSENTATION DU CORPUS

2.1. Les communications entre contrôleurs et pseudo-pilotes

Le corpus sur lequel porte notre étude est constitué d'enregistrements de dialogues oraux spontanés entre des contrôleurs aériens en formation et des « pseudos-pilotes ». Ces derniers sont des instructeurs simulant le rôle de pilotes en exercice. Deux langues sont utilisées : le français (majoritaire) et l'anglais. Le but des exercices enregistrés est d'entraîner les apprentis contrôleurs, et ensuite de les évaluer. Il s'agit pour eux de gérer plusieurs avions situés dans la zone contrôlée, par exemple en leur assignant une vitesse et/ou une position données. Pour des raisons techniques, le canal audio ne peut être « occupé » que par un seul locuteur à la fois, ce qui empêche tout recouvrement de parole.

Les productions orales des contrôleurs et des pilotes sont gouvernées par une stricte phraséologie, présentée dans [2]. Celle-ci décrit, par exemple, la manière dont les locuteurs doivent prononcer les identifiants des avions, ou bien l'ordre que doivent observer les différentes parties d'un message[2]. Durant la formation, ainsi d'ailleurs que dans des conditions réelles de travail, la phraséologie n'est pas toujours systématiquement appliquée. Le cadre général qu'elle fixe est cependant respecté.

[1] Plusieurs autres termes synonymes sont utilisés dans la littérature.

[2] Pour une description des indicatifs français et des ordres, se référer à Dourmap & Truillet [7].

Il est également important de noter que les dialogues enregistrés appartiennent bien à la catégorie du discours oral spontané. Nous tenons à le préciser car le rôle prépondérant tenu par la phraséologie pourrait laisser à penser que tous les énoncés produits sont déjà planifiés à l'avance. Or, ce n'est pas le cas : ni les contrôleurs, ni les pilotes ne savent à l'avance ce qui va arriver, et par conséquent ce qu'ils vont avoir à dire. La phraséologie définit seulement le cadre général de production des énoncés ; ce qui est dit repose sur l'interaction dynamique entre un contrôleur et pilote ou pseudo-pilote donnés, en fonction d'une situation variable.

2.2. Méthodologie de transcription et d'annotation

Nous avons procédé à la transcription et l'annotation des dialogues selon les spécifications de Coullon & Graglia [5].

Ces spécifications ont pour but de déterminer les éléments à transcrire, d'obtenir l'homogénéité des transcriptions dans le cas où plusieurs annotateurs se succèdent. Elles consistent essentiellement en règles à suivre pour transcrire les termes techniques tels que les indicatifs, les vitesses, etc. Elles donnent également des instructions de transcription des phénomènes tels que les hésitations ou les pauses. Nous avons ajouté à ces spécifications quelques autres classes et sous-classes de phénomènes devant être transcrits.

Le logiciel Transcriber[3] 1.4.2 a été utilisé pour les transcriptions.

2.3 Caractéristiques du corpus

Les enregistrements ont été effectués avec un DAT (Digital Audio Tape), et échantillonnés à 16 kHz (16 bits). Pour des raisons d'enregistrement, la qualité sonore souffre parfois de problèmes résultants de la saturation ou de bruits tels que les interférences ; cependant, les dialogues sont intelligibles. La table 1 ci-dessous présente les principales caractéristiques du corpus[4].

Table 1: Principales caractéristiques de notre corpus

Durée	Nombre de locuteurs	Nombre de tours de parole	Nombre de mots
36h50mn	16 (répartis en 2 groupes)	11 427	76 306

3. LES PHÉNOMÈNES DE DYSFLUENCE

3.1. Quelques points de terminologie

Dans la littérature, les termes utilisés par les auteurs pour désigner un phénomène donné varient souvent. Pour cette raison, nous présentons ci-dessous la terminologie (en français) employée dans notre travail.

Nous avons défini 6 différentes catégories de dysfluences. Lorsque cela est nécessaire, nous donnons des exemples pour illustrer notre propos (en mettant la dysfluence en gras).

- Hésitation : désigne l'interjection « euh ». Selon certaines terminologies (notamment Henry *et al.* [10]), il appartient à la catégorie des « pauses remplies ». Exemple:

maintenons niveau 1 0 0 Poitiers Amboise ***euh*** *Lacan*

- Répétitions : un mot (ou un groupe de mots) apparaît au moins deux fois à la suite. Nous n'avons pas pris en compte la répétition de dysfluences. Exemple:

station ***station*** *calling euh repeat your callsign*

- Amorce : l'arrêt de la production d'un mot avant la fin normale de celui-ci. Dans notre terminologie, une amorce correspond toujours à un fragment de mot que l'on peut identifier (souvent grâce à la connaissance de la phraséologie). Exemple (l'amorce est entre crochets) :

speed euh 200 Kts ***[mak]*** *euh minimum*

Le contexte et la phraséologie aident à comprendre que le contrôleur commence à prononcer « maximum ». Il se rend compte que cela ne convient et s'interrompt (« mak »). Enfin, il dit le mot correct : « minimum ».

- Fragment de mot : un ou plusieurs phonèmes inidentifiables (par opposition aux amorces). Exemple (le fragment est entre crochets) :

due to ***[ou]*** *due traffic euh descend level 9 0*

- Allongement : l'allongement d'une unité phonétique d'un mot, supérieur à 0,5 sec. Peut être combiné aux hésitations. Ce phénomène entre également dans la catégorie des « pauses remplies ».
- Pause longue : toute pause supérieure à 0,5 seconde et comprise à l'intérieur un tour de parole

3.2. Distribution des phénomènes

La figure 1 présente la répartition des catégories décrites ci-dessus. Les nombres situés immédiatement après le nom de la catégorie correspondent au nombre total d'occurrences relevées ; les pourcentages (en gras) sont calculés par rapport au nombre total d'occurrences de dysfluences.

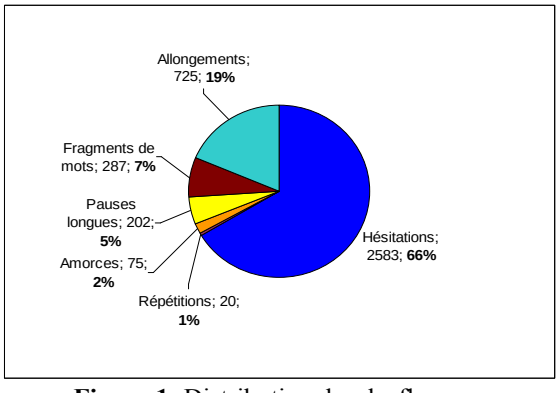

Figure 1: Distribution des dysfluences

[3] http://www.etca.fr/CTA/gip/Projets/Transcriber/IndexFr.html

[4] Le corpus (oral et transcription) n'est pas disponible sans demande préalable auprès de l'ENAC. S'adresser aux auteurs pour plus de renseignements, ou pour demander des échantillons.

Mettons cette distribution en perspective au moyen d'une comparaison détaillée avec d'autres études sur le même thème. Voici d'abord une courte description (nature de la tâche, nombre de mots, etc.) de chaque corpus d'oral spontané sur lesquels ces études sont basées :

- Candéa [4][5] : corpus de 13 histoires racontées oralement par des enfants. Durée : 70 minutes et 25 secondes ;
- Henry & Pallaud [9] : corpus de 1 000 382 (différentes situations d'oral spontané ; 794 locuteurs différents);
- Shriberg [13] : corpus de 54 minutes, et comprenant 8500 mots. Les 10 locuteurs parlent de leur travail ou de leur souvenirs ;
- Kurdi [11] : basée sur corpus de négociations (en Anglais) de transport de marchandises par train. Il comprend 52 000 mots.

On le voit, ces études sont basées sur des corpus très variés, que ce soit au niveau de la tâche ou de la taille. Cette diversité constitue une base pertinente de comparaison avec notre propre corpus.

Nous allons maintenant présenter les comparaisons pour chacune des catégories de dysfluences que nous avons définies. Evidemment, puisque chaque étude ne couvre pas l'ensemble des dysfluences, nous présenterons seulement celles qui concernent un phénomène donné, ou dont la catégorisation est proche de la nôtre[6].

- Répétitions : on voit par la table 2 que notre corpus comprend beaucoup moins de répétitions que les autres corpus. Nous pensons que la principale explication repose sur la nature même du corpus. Notre hypothèse est que, dans un contexte de contrôle aérien, les locuteurs (contrôleurs et pilotes) doivent être particulièrement vigilants afin d'éviter les ambiguïtés ou problèmes qui pourraient nuire à la compréhension de l'énoncé. De même, le temps nécessaire pour produire un énoncé n'est pas extensible : le locuteur ne peut pas perdre trop de temps en hésitation ou autres pauses (remplies ou silencieuses). Comme nous le verrons plus bas, cette hypothèse est confirmée par le fait que, pour chacune des autres catégories de dysfluences que nous avons définies, il y a systématiquement moins d'occurrences dans notre corpus (proportionnellement à la taille du corpus de comparaison) ;

Table 2: comparaisons pour les répétitions

Nom de l'étude	Notre corpus	[4]	[13]	[11]
Nombre de répétitions	20	110	141	256

- Hésitations : comme on le voit dans la table 3, il y a également bien moins d'hésitations dans notre corpus que dans ceux des autres travaux référencés. Ainsi, il y a 544 occurrences dans [4], mais ce corpus ne dure que 70 minutes, contre 35 heures pour le nôtre. De ce fait, il y a proportionnellement moins d'occurrences dans le corpus de [4]. On peut cependant noter que la différence semble globalement moindre que celle que nous avons observée pour les répétitions ;

Table 3: comparaisons pour les hésitations

Nom de l'étude	Notre corpus	[4]	[11]
Nombre et/ou pourcentage d'hésitations (par rapport au nombre total de mots)	2583 / **3.38%**	544	3512 / **6.75%**

- Amorces et fragments de mots : [9] est la seule étude dont la catégorisation est la plus proche de la nôtre en ce qui concerne les « amorces » et « fragments de mots ». Elle présente également des statistiques détailles sur leur distribution. Comme les auteurs ne font pas la distinction entre les « amorces » et les « fragments de mots », nous additionnerons les occurrences des deux types de phénomènes qui apparaissent dans notre corpus. Il en résulte un total de 362 occurrences, soit 0.47% du nombre total de mots. [9] fait état d'un total de 6094 occurrences des « fragments de mots » pour environ un million de mots (soit approximativement 0.6%). La distribution dans notre corpus de cette double catégorie est assez proche de celle observée dans [9], contrairement à ce que nous avons constaté pour les catégories précédentes. Toutefois, une explication à ce résultat pourrait être le fait que notre double catégorie ne correspond pas exactement à celle définie par [9] ;

Table 4: comparaisons pour les allongements

Nom de l'étude	Notre corpus	[4]	[13]
Nombre et/ou pourcentage d'allongements (par rapport au nombre total de mots)	725 / **0.9%**	284	669 (y compris "euh") / **7.9%**

- Allongements : une fois encore, comme on le voit dans la table 4, la fréquence de ce que nous appelons allongement est moindre dans notre corpus que dans les autres ;
- Pauses longues : la table 5 montre que la spécificité de notre corpus est un peu moins prononcée que pour les autres catégories de dysfluences. Cependant, là encore, on remarquera qu'il y a moins de " pauses longues" que dans d'autres corpus.

Table 5: comparaisons pour les pauses longues

Nom de l'étude	Notre corpus	[4]	[13]
Nombre et/ou pourcentage of pauses longues (par rapport au nombre total de mots)	827 / **1.08%**	1471	318 / **3.74%**

4. LE CAS DES AMORCES

Plusieurs intérêts scientifiques nous poussent à nous pencher sur le cas des amorces, bien qu'elles soient peu représentées dans notre corpus. Le principal est qu'elles permettent, dans de nombreux cas, d'avoir une idée de la planification de la production du locuteur, comme nous allons le montrer ci-après.

Plus particulièrement, nous nous sommes intéressés aux amorces correspondant à une auto-correction du locuteur : celles où l'arrêt en cours de production caractéristique des

[5] Pour plus de lisibilité, nous désignons dans la suite de l'article les études uniquement par leur numéro de référence.

[6] Lorsque cela est possible, nous indiquons dans les tableaux le nombre d'occurrences et le pourcentage. Nous mettons en gras cette dernière mesure, afin de faciliter la lecture.

amorces est motivé par la prise de conscience du locuteur qu'il fait une erreur[7]. Nous avons distingué quatre sous-catégories, en fonction précisément de la nature de l'erreur qui provoque l'interruption de la production :

- Erreur sur un « mot » : nous appelons « mot » les données alpha-numériques (indicatifs, par exemple), et les commandes (« grimpez », « demande », etc.) Exemple :

climbing for level 1 7 0 and 2 0 0 Kts ***[mak]*** *euh minimum D M C*

- Erreur sur l'organisation de l'énoncé : un mot ou un groupe de mots n'occupant pas sa position correcte dans l'énoncé.

[poi] *Absie Poitiers Balon Reson Britair B X*

- Erreur sur la langue utilisée : le locuteur remarque (ou bien on lui fait remarquer) qu'il n'a pas parlé dans la langue appropriée : français à la place de l'anglais ou vice versa.

P I ***[vite]*** *speed 2 1 0 Kts*

- Erreur de prononciation : comme son nom l'indique…

c'est le ***[lio]*** *Littoral*

La distribution de ces différentes sous-catégories est donnée dans la figure 2 ci-dessous.

On constate que la majorité des erreurs concerne ce que nous avons appelé « mot ». Nous attribuons cela à la charge cognitive élevée que cette catégorie peut induire. En effet, il s'agit très souvent de termes que les apprentis contrôleurs ne sont pas encore habitués à « manier ». La forte charge cognitive ainsi engendrée est elle-même la source de problèmes de production.

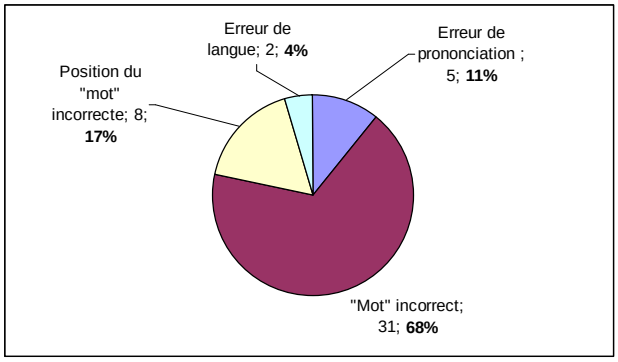

Figure 2 : Distribution des amorces correctives

5. CONCLUSION

Notre corpus présente un certain nombre de caractéristiques différentes de celles de corpus plus « traditionnels », par exemple de demandes d'informations. Nous avons montré qu'il y a une très forte différence entre les fréquences d'apparition de dysfluences dans notre corpus par rapport à d'autres corpus, notamment en ce qui concerne les répétitions. L'explication plausible de ce phénomène repose sur la spécificité de ce corpus. En nous penchant sur le cas plus particulier des « amorces », nous avons également mis en évidence la distribution particulière de cet autre type de phénomènes, que nous avons attribuée à la spécificité de la tâche.

De nombreuses prolongations à cette première étude sont envisagées, telles que la caractérisation acoustique (intensité, durée, fréquence fondamentale, représentation spectrale) des dysfluences, pour améliorer leur reconnaissance au niveau du décodage acoustique.

Une caractérisation plus fine des propriétés linguistiques du corpus dans la perspective d'une modélisation stochastique dans le cadre d'un module de compréhension de la parole est en cours. Elle nous permettra de procéder à l'implémentation du modèle obtenu.

Enfin, nous poursuivons un autre axe de recherche : l'analyse des rapports entre la situation d'apprentissage, la charge cognitive des locuteurs, et les performances de ceux-ci.

Nous mettons actuellement en place en place un protocole d'évaluation reproduisant des conditions de contrôle aérien aussi proche que possible de la réalité. Il nous servira à mettre en œuvre les deux dernières perspectives.

BIBLIOGRAPHIE

[1] M. Adda-Decker, B. Habert, C. Barras, G. Adda, P. Boula De Mareuil, P. Paroubek. Une étude des dysfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. *(JEP'04)*. 2004.

[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.

[4] M. Candéa. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané.* Thèse d'État, Université Paris III (Sorbonne Nouvelle), 2000.

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

[7] L. Dourmap & P. Truillet. *Interaction vocale dans le contrôle aérien : la comparaison de deux grammaires contextuelles pour la reconnaissance des indicatifs de vol.* CENA internal report. 2003.

[8] P. Falzon *Ergonomie cognitive du dialogue*. Presses Universitaires de Grenoble, 1989.

[9] S. Henry & B. Pallaud. Word fragments and repeats in spontaneous spoken French, *DiSS'03,* 2003.

[10] S. Henry, E. Campione & J. Véronis. Répétitions et pauses (silencieuses et remplies) en français spontané. *(JEP'04)*. 2004.

[11] M.- Z. Kurdi. *Contribution à l'analyse du langage oral spontané.* Thèse de doctorat, Université J. Fourier, Grenoble, France, 2003.

[12] P. Lerat. *Les langues spécialisées.* Paris ,PUF, 1995.

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

[7] Cela concerne 29 amorces, soit 39% de leur nombre total.